U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

Modelação e Previsão de Anulações no Seguro Automóvel

Beatriz Sousa Faustino

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto
Mestrado em Engenharia Matemática

2016

MSc

2.º CICLO

FCUP
2016

U.PORTO

Modelação e Previsão de Anulações no Seguro Automóvel

Beatriz Sousa Faustino

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

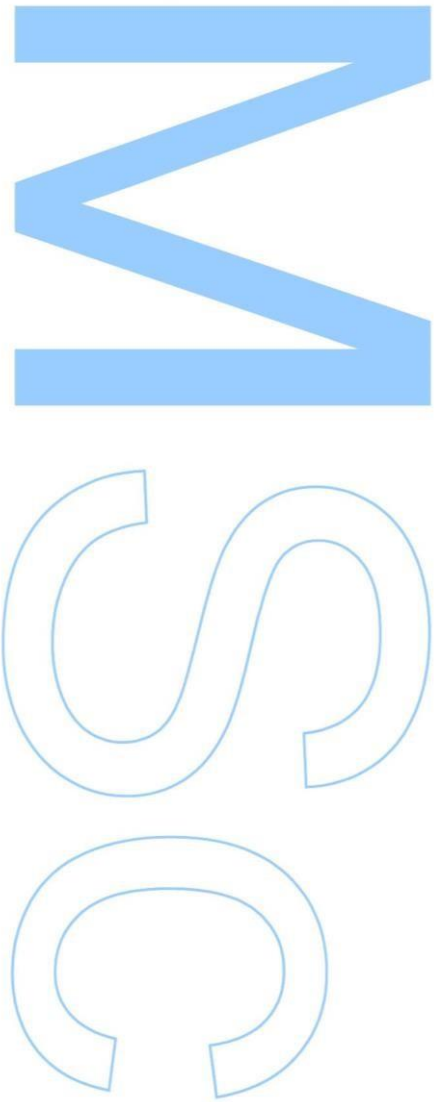
MSc

Modelação e Previsão de Anulações no Seguro Automóvel

Beatriz Sousa Faustino
Mestrado em Engenharia Matemática
Departamento de Matemática
2016

Orientador
Professor Doutor Joaquim Fernando Pinto da Costa, FCUP

Orientador de Estágio
Dr. Luís Maranhão, Ageas Portugal




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

Agradeço ao Professor Doutor Joaquim Fernando Pinto da Costa e ao Dr. Luís Maranhão pela orientação nesta dissertação, a qual não poderia ter sido realizada sem os conselhos, a paciência, disponibilidade e atenção mostrada por ambos.

Agradeço à AXA Portugal e à Ageas Portugal por me terem proporcionado este estágio curricular de 9 meses, o qual me permitiu aplicar e adquirir novos conhecimentos.

Agradeço também à minha família por todo o apoio disponibilizado. Em especial à minha mãe e ao meu irmão, pelo incentivo e apoio incondicional na minha formação académica que permitiu que esta jornada acontecesse.

Por fim, agradeço aos meus colegas da AGEAS, Luís Filipe Ferreira, Filipe Gonçalves, Joana Garraio, Andreia Ginja e Tiago Morais pela amizade, ajuda e conselhos.

Resumo

A concorrência na atividade seguradora tem sido cada vez maior e por essa razão, é importante poder prever e compreender o comportamento dos clientes.
Este trabalho teve como objetivo a construção de um modelo de regressão logística a fim de prever possíveis anulações de contratos no seguro automóvel por parte do cliente, tendo sido para isso recolhidas várias variáveis que pudessem ter influência na decisão de anulação do contrato.
Foi feita inicialmente uma análise descritiva dessas variáveis para perceber como cada uma delas afeta individualmente a resposta (se o cliente anula ou não). De seguida, foram eliminadas as variáveis redundantes, foram agrupadas algumas categorias com frequências próximas de zero em algumas das variáveis e os dados foram divididos num conjunto de treino e de teste antes de ser feito o ajuste do modelo de regressão logística.
Com vista a escolher o melhor modelo, foram ajustados vários modelos. Foram considerados no total 17 modelos e para que fosse possível a comparação e escolha do melhor modelo recorreu-se às medidas de AIC e AUC.
Por fim, mediu-se a capacidade preditiva do modelo final ajustado.

Palavras-chave: regressão logística, anulações no seguro automóvel, dados categóricos.

Abstract

Competition in the insurance business has been growing and therefore it is important to be able to anticipate and understand customer behavior.

This work aimed to build a logistic regression model in order to predict possible contract cancellations in the field of auto insurance by the customer, having collected several variables that could influence the contract annulment decision.

Initially a descriptive analysis of these variables was made to see how each variable affects the response (if the customer cancels or not). Then redundant variables were eliminated , some categories with near zero frequencies in some of the variables were grouped and the dataset was divided into a training and a test set before making the adjustment of the logistic regression model.

In order to choose the best model, several models were adjusted. In total 17 models were considered and to be able to compare and choose the best one among these we used AIC and AUC measures.

Finally, we measured the predictive ability of the final adjusted model.

Keywords: logistic regression, annulments in auto insurance, categorical data.

Conteúdo

Lista de Quadros

Lista de Figuras

Capítulo 1

Apresentação da empresa

Este estágio curricular teve a duração de 9 meses, com início em 5 de Outubro de 2015 e fim em 4 de Julho de 2016.

Até 1 de Abril o estágio foi realizado na AXA Portugal tendo esta sido posteriormente vendida à empresa Ageas.

A história da AXA é rica e remonta a 1835. Enquanto marca, a AXA celebrou em 2010 os 25 anos de existência a nível mundial. Em 2013, foi pelo 5.º ano consecutivo, a melhor marca mundial de seguros (Best Global Brands – Top 100 – Interbrand).

No âmbito da Proteção Financeira, o negócio da AXA Portugal consistia na oferta de soluções para os ramos Vida e Não Vida. O ramo Vida engloba seguros de vida, soluções de reforma e outros investimentos, enquanto o ramo Não Vida abrange seguros pessoais e patrimoniais.

Dia 26 de Abril foi lançada oficialmente em Portugal a nova marca Ageas Seguros que veio substituir a AXA Portugal.

A Ageas é um grupo segurador internacional, sediado em Bruxelas, com 190 anos de experiencia. Presente em 13 países da Europa e da Asia, a empresa propõe soluções de seguros de vida e não vida a milhões de Clientes Particulares e Empresas.

A Ageas é um dos maiores grupos seguradores europeus, é líder na Bélgica e encontra-se entre os principais “players” (intervenientes) na maioria dos países em que está presente. Está presente em Portugal desde 2005, operando já através de marcas reconhecidas, como a Médis e a Ocidental.

Figura 1. 1 - Países em que a Ageas se encontra presente

Capítulo 2

Introdução

2.1 Breve introdução ao contrato de seguro

A segurança do homem é permanentemente sujeita a inúmeras ameaças e perigos que põem em causa a sua integridade física ou patrimonial. Isto leva a que se tomem determinadas medidas de prevenção e de proteção que visam a sua salvaguarda e os seguros são uma das medidas mais eficazes para tal. Representam pois formas avançadas de proteger património e investimentos, e contribuem, decisivamente, para o desenvolvimento das sociedades.

O seguro é uma operação que toma forma jurídica de um contrato em que o segurador assume a cobertura de determinados riscos do tomador de seguro ou de outrem, comprometendo-se a satisfazer a prestação convencionada em caso de ocorrência de sinistro. Em contrapartida, a pessoa ou entidade que celebra o seguro (tomador de seguro) fica obrigada a pagar ao segurador o prémio correspondente, ou seja, o custo do seguro.

O contrato de seguro é sempre feito segundo o interesse de alguém e essa pessoa é a segurada. Se é simultaneamente tomador assume os mesmos direitos e obrigações que esta. No caso do seguro automóvel, a pessoa segurada é o condutor do veículo.

A apólice de seguro é o documento que titula o contrato celebrado entre o Tomador de Seguro e o Segurador, onde constem as respetivas condições gerais, especiais, se as houver, e particulares acordadas.

2.2 Objetivo deste trabalho

A duração de um contrato de seguro é o período dentro do qual vigora o mesmo. Este período pode ser determinado (seguro temporário) ou pode ser renovado anualmente, mensalmente, semestralmente ou trimestralmente, exceto se qualquer das partes denunciar o contrato com 30 dias de antecedência mínima em relação à data de

vencimento (data em que o contrato termina) ou se o tomador do seguro não proceder ao pagamento do prémio.

O objetivo deste estudo consistiu na identificação de fatores de risco associados à anulação do contrato de seguro automóvel por parte do cliente.

Tendo em atenção o objetivo do estudo, todas as apólices temporárias foram removidas.

Os dados utilizados nesta análise contêm informação de Novembro de 2009 a Dezembro de 2015.

Um contrato é considerado anulado se a data de anulação ocorrer no período de 60 dias antes da data de vencimento até 90 dias depois da data de vencimento. Desta maneira, para ver, por exemplo, se um contrato com data de vencimento em Janeiro de 2010 foi ou não anulado, é necessário analisar as anulações dos meses de Novembro e Dezembro de 2009 e dos meses de Fevereiro, Março e Abril de 2010. Assim os vencimentos considerados em 2015 foram apenas os relativos aos 9 primeiros meses desse ano para que passados 90 dias (Dezembro) ainda se pudesse verificar as anulações relativas a Setembro de 2015.

2.3 Softwares utilizados

Os softwares utilizados, tanto na análise descritiva como na construção do modelo de regressão logística foram o SAS e o Emblem.

2.4 Variáveis em estudo

As variáveis recolhidas para este estudo foram as seguintes:

Variável	Descrição	Valores
Codanula	Indica se o contrato se encontra em vigor ou se foi anulado.	0 – Em vigor
		1 – Anulado
Sinistralidade	A variável Sinistralidade indica se o cliente teve ou não sinistros nos últimos 5 anos.	Não teve sinistros
		Teve Sinistros
Semestre do Vencimento	É o semestre em que se dá o vencimento do contrato de seguro.	1º Semestre
		2º Semestre
Outras Apólices	Se o cliente tem ou não outras apólices em vigor na seguradora.	Sim
		Não

Quadro 2. 1 - Variáveis em estudo e as suas respectivas descrições e valores

Variável	Descrição	Valores
Outras Apólices Não Vida	Se o cliente tem ou não outras apólices em vigor pertencentes ao ramo não vida na seguradora.	Sim
		Não
Outras Apólices Automóvel	Se o cliente tem ou não outras apólices em vigor pertencentes ao ramo automóvel na seguradora.	Sim
		Não
Antiguidade do Contrato	A antiguidade do contrato é o número de anos de vigência do contrato de seguro automóvel.	≤ 1 ano
		]1,2] anos
		[3,4] anos
		[5,10] anos
		> 10 anos
Escalão de Bónus	O sistema bónus-malus consiste numa escala progressiva (0 a 26) de agravamentos por sinistro e de bónus por ausência do mesmo. Se o cliente tem escalão entre 0 e 6 então o prémio sofre um agravamento, caso contrário, sofre um desconto. O agrupamento à direita é feito com base nas tarifas automóveis.	1 se $\in [0,6]$
		2 se =7
		3 se = 8
		4 se $\in [9,10]$
		5 se $\in [11,13]$
		6 se $\in [14,17]$
		7 se $\in [18,26]$
Evolução Bónus-Malus	Corresponde à evolução no escalão de bónus.	Evolução Negativa
		Sem Evolução
		Evolução Positiva
Sexo	Sexo do tomador do seguro automóvel.	Jurídico (se o cliente é empresa)
		Feminino
		Masculino

Quadro 2.1 – Continuação da página anterior

Variável	Descrição	Valores
Intervalo estimado de variação do prémio	A atribuição de bónus/agravamento ao valor do prémio depende do nº de anos de vigência do contrato e do nº de sinistros registados. Assim, o escalão de bónus de um cliente tem que ser revisto anualmente, diminuindo-o ou aumentando-o consoante ocorra ou não sinistros respetivamente. Antes de alterar o valor do prémio, o segurador deve avisar o tomador de seguro trinta dias antes da data de vencimento.	Redução> 50€
		Redução ∈]25€;50€]
		Redução ∈]20€;25€]
		Redução ∈]15€;20€]
		Redução ∈]10€;15€]
		Redução ∈]5€;10€]
		Redução ∈]0€;5€]
		0 €
		Aumento ∈]0€;4.5€]
		Aumento ∈]4.5€;10€]
		Aumento ∈]10€;15€]
		Aumento ∈]15€;20€]
		Aumento ∈]20€;50€]
		Aumento ∈]50€;100€]
		Aumento> 100€
Região	Região onde reside o tomador do seguro automóvel.	Entre Douro e Minho
		Trás-os-Montes e Alto
		Grande Porto
		Beira Litoral
		Beira Interior
		Estremadura e Ribatejo
		Lisboa
		Setúbal
		Alentejo
		Algarve
		Ilhas

Quadro 2.1 – Continuação da página anterior

Variável	Descrição	Valores
Forma de Cobrança	Meio de pagamento do prémio do seguro automóvel.	Agente cobrador
		Banco – DACB
		Outros
		Tesourarias
Forma de Pagamento	Intervalo de tempo até efetuar novo pagamento do prémio.	Anual
		Semestral
		Trimestral
		Mensal
Categoria Automóvel	Categoria do automóvel seguro.	Ligeiros
		Motociclos
		Outros
		Pesados
Idade do Tomador	Idade do tomador do seguro automóvel.	Até aos 20 anos
		Dos 21 aos 25 anos
		Dos 26 aos 30 anos
		Dos 31 aos 40 anos
		Mais de 40 anos
		NA (se o cliente é empresa)
Idade do Condutor	Idade do condutor do veículo seguro.	Até aos 20 anos
		Dos 21 aos 25 anos
		Dos 26 aos 30 anos
		Dos 31 aos 40 anos
		Mais de 40 anos
		NA (se o cliente é empresa)
Tempo de Carta	Tempo de carta do condutor do veículo seguro.	0 anos
		1 a 2 anos
		3 a 5 anos
		6 a 10 anos
		Mais de 10 anos
		NA (se o cliente é empresa)
Rede	Meio através do qual o cliente realizou o seu contrato de seguro automóvel.	Lojas
		Outros
		Pontos
		Private
		RNA
		Unidade Bancárias e Parcerias

Quadro 2.1 – Continuação da página anterior

Variável	Descrição	Valores
Pack Automóvel	O seguro de responsabilidade civil é um seguro obrigatório no que respeita à cobertura de responsabilidade civil. Este seguro garante em caso de acidente a reparação dos danos materiais ou corporais causados a terceiros. Por outro lado existe o seguro de danos próprios que abrange os prejuízos sofridos pelo veículo seguro independentemente de quem é responsável pelo acidente, mesmo que seja o próprio condutor. Este ao contrário do seguro de responsabilidade civil, é um seguro facultativo.	Responsabilidade Civil
		Danos Próprios

Quadro 2.1 – Continuação da página anterior

Variável	Descrição	Valores
Produto	Tipos de seguros existentes na companhia.	Protect
		Ice 3
		Protocolos-Ordens
		Protocolos-Renault
		Protocolos Financeiras
		Protocolos-Barclays
		Protocolos-FSegurança
		Protocolos-AssComerciais
		Protocolos-Sonae
		Protocolos-Outros
		Protocolos-Funcionários
		Protocolos-Seguros Especiais

Quadro 2.1 – Continuação da página anterior

2.5 Análise descritiva das variáveis

Os dados descritos e o problema colocado inserem-se num contexto de regressão logística: a resposta é a variável Codanula, sendo as restantes as variáveis explicativas. Todas as variáveis são categóricas dicotómicas ou policotómicas.
Porém, antes de iniciar a análise, tentou-se perceber de que forma é que os dados se estão a comportar, analisando a relação entre a resposta e as várias variáveis explicativas a considerar no modelo de regressão. Para isso observaram-se os gráficos de percentagens de anulações por valor de cada variável e realizou-se o teste de independência entre dois fatores (entre a resposta e cada variável explicativa) [1].

2.5.1 Gráficos de percentagens de anulações de apólices por valor de cada variável

É de notar que os gráficos seguintes avaliam apenas a relação bruta da variável em causa com a resposta. No modelo de regressão a relação a avaliar será a associação ajustada para as restantes variáveis, que claro pode ser diferente da primeira.
De forma a não revelar informação confidencial da Seguradora, nos gráficos que se seguem, a percentagem de anulações por categoria da variável é calculada como sendo a percentagem de anulações dessa mesma categoria sobre a percentagem de anulações da categoria com maior número de anulações. Ou seja, considerando uma variável qualquer, vê-se as percentagens de anulações de cada categoria e toma-se como referência, aquela com maior percentagem de anulações.

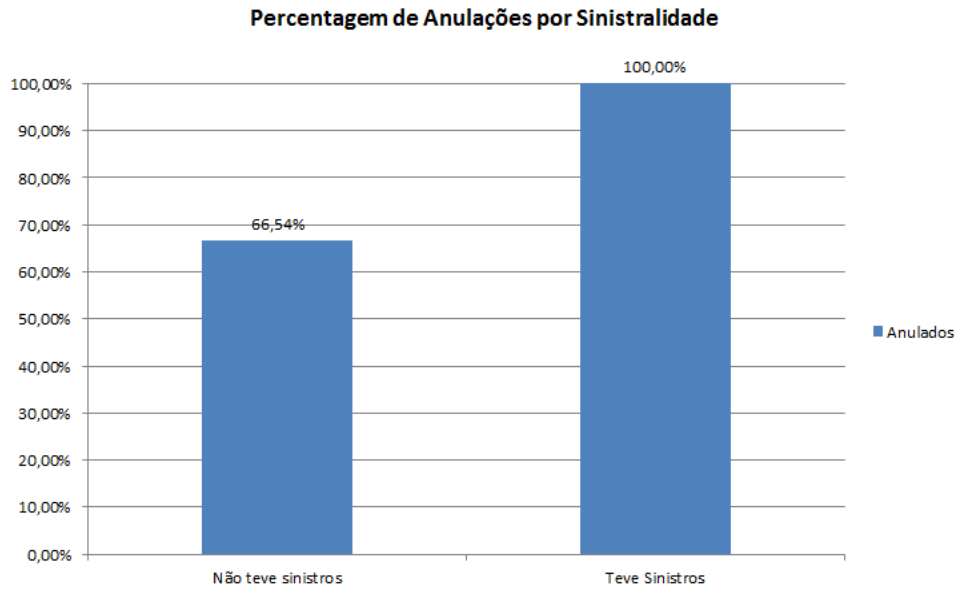


Figura 2. 1 - Anulações vistas pelos valores da variável Sinistralidade

Da análise da figura 2.1 conclui-se que a percentagem de anulações na população de clientes que tiveram sinistros nos últimos 5 anos é superior à percentagem de anulações na população de clientes que não tiveram sinistros nos últimos 5 anos.
Esta observação já era esperada uma vez que a ocorrência de sinistros faz baixar o valor no escalão de bónus do cliente, aumentando desta maneira o valor do prémio.

De maneira semelhante foram feitas as análises dos gráficos de percentagem de anulações para as restantes variáveis, encontrando-se estes gráficos na parte dos Anexos (Anexo A).

Figura A.1: Observa-se que quanto maior é o tempo de contrato menor é a percentagem de anulações. No entanto, para antiguidades do contrato ≤ 1 ano, 2 anos e de 3 a 4 anos, a percentagem de anulações não difere muito.

Figura A.2: A percentagem de anulações diminui à medida que aumenta o escalão de bónus. Isto deve-se ao facto de que um valor acima no escalão de bónus corresponde a um desconto maior no prémio, o valor 1 do escalão de bónus corresponde a agravamentos e portanto é o que tem maior percentagem de anulações.

Figura A.3: As populações de clientes que sofreram aumentos no valor do prémio têm percentagens de anulações superiores às das populações de clientes que viram o seu prémio reduzido.
É de notar, no entanto, que as populações de clientes que usufruíram de uma redução $\in$]20€;25€], $\in$]5€;10€], $\in$]0€;5€] ou de 0€ (sem variação no valor do prémio) têm percentagens de anulações semelhantes. Também os clientes que obtiveram uma redução $\in$]15€;20€] e uma redução $\in$]10€;15€] têm percentagens de anulações próximas entre si e maiores do que por exemplo a população de clientes que não teve qualquer alteração no valor do prémio.
Com menor percentagem de anulações estão aqueles clientes que viram o seu prémio reduzido num valor >50€.

Figura A.4: O semestre no qual se dá o vencimento parece não apresentar qualquer influência sobre a percentagem de anulações do contrato de seguro.

Figura A.5: A população de clientes que evoluiu de forma positiva no escalão de bónus é a que apresenta maior percentagem de anulação comparativamente com as populações de clientes que evoluíram negativamente ou que não tiveram evolução. Isto pode ser devido ao facto de a evolução no escalão de bónus corresponder a uma redução pequena no valor do prémio e portanto, não corresponder à expectativa do cliente.

Figura A.6: Pode-se concluir que na população de clientes que têm outras apólices em vigor na seguradora, além da apólice automóvel em estudo, a percentagem de anulações é muito menor comparada à percentagem de anulações da população de clientes que só têm uma apólice na seguradora.

Figura A.7: Especificando o tipo de outras apólices que o tomador do seguro possa ter na seguradora, tem-se que se o cliente tiver outras apólices do ramo não vida então a tendência mantém-se e a percentagem de anulação é menor na população de clientes que têm outras apólices em vigor no ramo não vida.

Figura A.8: Para os clientes que têm outras apólices no ramo automóvel a percentagem de anulações é também menor comparando com a percentagem de anulações dos clientes que só têm uma apólice na seguradora.

Figura A.9: As percentagens de anulações por região em geral não diferem muito. Com menor percentagem de anulações tem-se a região de Trás-os-Montes e Alto e com maior percentagem de anulações tem-se as Ilhas.

Figura A.10: Os veículos pesados normalmente são contratualizados por empresas o que poderá explicar a maior percentagem de anulações comparada com as percentagens de anulações das outras categorias automóvel. Com menor percentagem de anulações encontra-se a categoria outros (outros veículos além dos ligeiros, motociclos e pesados).

Figura A.11: Não há grande diferença entre a percentagem de anulações da população de clientes que possui Responsabilidade Civil e a percentagem de anulações da população de clientes que possui Danos Próprios.

Figura A.12: A maior percentagem de anulações encontra-se na população de clientes que adquirem o produto Protocolos-Renault e a menor percentagem de anulações encontra-se na população de clientes que têm o produto Protocolos-Funcionários (produto destinado a funcionários da companhia seguradora).

Figura A.13: A população de clientes que paga anualmente o valor do prémio tem maior tendência a anular do que as populações de clientes que pagam o prémio semestralmente, mensalmente e trimestralmente.

Isto é explicado pelo facto de que se houver um aumento no valor do prémio para apólices que são renovadas anualmente então esse aumento vai ser pago de uma só vez, enquanto que para apólices que são renovadas semestralmente, mensalmente ou trimestralmente esse aumento vai ser distribuído por períodos de tempo mais curtos o que o torna menos percetível.

Figura A.14: A população de clientes que paga o valor do prémio em tesourarias é a que tem maior percentagem de anulações. Logo a seguir, está a população de clientes que paga o valor do prémio através de agentes cobrador.
Com menor percentagens de anulações estão as populações de clientes que pagam por Bancos e outros meios.
Assim a tendência é a de que os clientes que pagam por débito directo têm menor sensibilidade ao valor do prémio que lhes é cobrado.

Figura A.15: Se o contrato de seguro automóvel for realizado em unidades bancárias e parcerias então é mais provável que o cliente anule o contrato.

Figura A.16: As populações de clientes femininos e masculinos não apresentam diferença a nível de percentagens de anulações.
Já no caso em que o cliente é empresa, a percentagem de anulações sobe um pouco.

Figura A.17: É de notar que a população de tomadores do seguro com mais de 40 anos é a que apresenta menor percentagem de anulações, contrariamente à população de tomadores do seguro com idade até aos 20 anos, que detém a maior percentagem de anulações.

Figura A.18: As conclusões aqui retiradas para a variável idade do condutor são as mesmas da análise da figura A.17 uma vez que a grande maioria dos tomadores do seguro também são condutores do veículo seguro. Note-se também que quanto maior é a idade, maior tende a ser o número de anos de vigência do contrato e se o número de sinistros for relativamente baixo, então o valor do prémio do seguro vai ser menor do que para aqueles que têm menor tempo de contrato.

Figura A.19: A população de clientes com tempo de carta superior a 10 anos é a que detém menor percentagem de anulações.

2.5.2 Teste de independência entre dois fatores, a resposta e cada variável

Os seguintes dados são necessários para a realização do teste de independência entre dois factores [2]:

- 1 População
- 2 Fatores (A e B)
- R níveis distintos do fator A
- C níveis distintos do fator B
- $O_{i,j} = \#$ indivíduos observados com factor A no nível i e factor B no nível j

O teste de hipóteses para a independência entre dois fatores, realiza-se sob as seguintes hipóteses:

H_0 : factor A independente do factor B

H_1 : os dois factores não são independentes

Estatística de teste:

$$U = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \dot{\sim} \chi^2\big((R-1)(C-1)\big) \tag{2.1}$$

Válida apenas se $E_{i,j} \geq 5, \forall i, j$ (todas as contagens esperadas são ≥ 5)

A decisão consiste em rejeitar H_0 com nível de significância α se:

$$U > \chi^2{}_{1-\alpha}\big((R-1)(C-1)\big) \tag{2.2}$$

Como todas as contagens esperadas são ≥ 5 em todas as células das diferentes tabelas de contingência entre a resposta e cada uma das variáveis então pode-se aplicar o teste de independência.

Na tabela seguinte seguem-se então os resultados para o Teste de independência entre a resposta e cada uma das variáveis:

Teste de independência entre a Resposta e cada uma das Variáveis		
Variável	**Estatística de Teste U**	**Valor p**
Sinistralidade	12221.3119	<0.0001
Antiguidade do Contrato	15701.2103	<0.0001
Escalão de Bónus	26484.2603	<0.0001
Intervalo Estimado de Variação do Prémio	32773.6478	<0.0001
Semestre do Vencimento	0.9616	0.3268
Evolução Bónus-Malus	6578.7113	<0.0001
Outras Apólices	178577	<0.0001
Outras Apólices Não Vida	24585.7681	<0.0001
Outras Apólices Automóvel	131824	<0.0001
Região	3686.2204	<0.0001
Categoria Automóvel	4718.3391	<0.0001
Pack Automóvel	216.1817	<0.0001
Produto	9105.9622	<0.0001
Forma de Pagamento	6164.7600	<0.0001
Forma de Cobrança	13908.9930	<0.0001
Rede	4430.0656	<0.0001
Sexo	2957.0365	<0.0001
Idade do Tomador	12375.3914	<0.0001
Idade do Condutor	12024.8288	<0.0001
Tempo de Carta	7360.7787	<0.0001

Quadro 2. 2 - Teste de independência entre a resposta e cada uma das variáveis

Exceto para a variável Semestre do Vencimento, as estatísticas de teste U e o valor-p do teste permitem rejeitar H_0 com um nível de significância de 0.0001 e concluir que cada uma das variáveis explicativas é dependente da resposta.

No caso da variável Semestre do Vencimento a estatística de teste obtida foi U=0.9616 e valor p=0.3268. Assim, com um nível de significância de 0.0001 não se rejeita H_0 e portanto não se pode concluir que haja dependência entre a resposta e a variável Semestre de Vencimento.

Claro que, devido a estarmos a realizar um grande número de testes, corremos sempre o risco de tirar conclusões erradas. No entanto, dado que o valor da prova de

cada um dos testes é altamente significativo, as conclusões são válidas mesmo se aplicarmos a correção de Bonferroni.

Capítulo 3

Regressão Logística

3.1 Odds e odds ratio

Todos os modelos estatísticos são representações abstratas e simplificadas da realidade.

Antes de explicar o conceito de regressão logística é importante compreender as definições de odds e de odds ratio.

O odds é definido como o quociente entre a probabilidade de um evento acontecer e a probabilidade desse mesmo evento não acontecer [1] e [3]:

$$\text{odds para evento} = \frac{\mathrm{P}(\text{evento})}{1 - \mathrm{P}(\text{evento})} \tag{3.1}$$

Já o odds ratio (OR) é definido como:

$$\mathrm{OR} = \frac{\dfrac{\mathrm{P}(\mathrm{A}=1|\mathrm{B})}{1-\mathrm{P}(\mathrm{A}=1|\mathrm{B})}}{\dfrac{\mathrm{P}(\mathrm{A}=1|\overline{\mathrm{B}})}{1-\mathrm{P}(\mathrm{A}=1|\overline{\mathrm{B}})}} = \frac{\dfrac{\mathrm{P}(\mathrm{A}=1|\mathrm{B})}{\mathrm{P}(\mathrm{A}=0|\mathrm{B})}}{\dfrac{\mathrm{P}(\mathrm{A}=1|\overline{\mathrm{B}})}{\mathrm{P}(\mathrm{A}=0|\overline{\mathrm{B}})}} \tag{3.2}$$

É portanto uma medida de associação entre uma exposição (B) e um resultado (A). O OR representa o odds de um evento ocorrer dada uma exposição particular, em comparação com o odds de o evento ocorrer na ausência dessa exposição. Ou seja, é o quociente entre esses dois odds.

O odds ratio pode tomar qualquer valor não negativo.

Quando o evento e a exposição são independentes, tem-se OR=1 [4].

Se OR>1 então o odds de um evento ocorrer dada uma exposição particular é superior ao odds de o evento ocorrer na ausência dessa exposição. É mais provável que o evento ocorra na presença da exposição referida do que na ausência dessa exposição. Dizemos que temos uma associação positiva entre o evento e a exposição.

Se OR<1 então o odds de um evento ocorrer dada uma exposição particular é inferior ao odds de o evento ocorrer na ausência dessa exposição. É mais provável que o evento ocorra na ausência da exposição referida do que na presença dessa

exposição. Dizemos que temos uma associação negativa entre o evento e a exposição.

Quanto mais afastado de 1 estiver o valor do OR, maior é a associação (negativa ou positiva) entre o evento e a exposição e portanto mais afastados de serem independentes estão.

3.2 Modelo de Regressão Logística

Tendo assimilado estas definições pode-se passar agora à compreensão do modelo de regressão logística.

O modelo de regressão logística é um caso particular dos modelos lineares generalizados com resposta binomial [1]. Assim existem apenas duas respostas possíveis, sucesso ou insucesso, ou seja,

Y=1, se o cliente anula o contrato de seguro

Y=0, se o cliente não anula o contrato de seguro

Como todos os modelos de regressão, o modelo de regressão logística tem como objetivo conseguir explicar ou prever a resposta binomial Y à custa de um conjunto de variáveis explicativas $X = (X_1, \dots, X_p)$.

Seja $x_i = (x_{i1}, \dots, x_{ip})$,uma observação do indivíduo i. Tem-se então que,

$$Y|X = x_i \sim B(1, \pi(x_i)) \tag{3.3}$$

onde,

$$\pi(x_i) = \pi_i = P(Y = 1|X = x_i) \tag{3.4}$$

é a probabilidade de sucesso para Y (probabilidade de o contrato ser anulado) dada a observação x_i. Pelas propriedades da distribuição binomial, a média de Y dado $X = x_i$ é:

$$\mu = E(Y|X = x_i) = \pi(x_i) \tag{3.5}$$

E a variância é:

$$Var(Y|X = x_i) = \pi(x_i)\big(1 - \pi(x_i)\big) \tag{3.6}$$

Note-se que o cálculo da média e da variância depende dos valores das variáveis explicativas e por esse motivo um modelo linear gaussiano não se adaptaria bem a esta situação uma vez que este último assume que a variância da resposta é constante.

Suponha-se agora que as observações sob estudo podem ser classificadas de acordo com as variáveis explicativas em k grupos de tal maneira que todos os indivíduos num

grupo têm valores idênticos em termos de mesmas categorias das variáveis explicativas. Estes grupos designam-se por padrões de covariáveis [5].

Por exemplo, suponhamos que o modelo tinha apenas duas variáveis explicativas: Sinistralidade e Categoria Automóvel.

	Ligeiros	Motociclos	Outros	Pesados
Não teve sinistros	$\frac{y_1}{n_1}$	$\frac{y_2}{n_2}$	$\frac{y_3}{n_3}$	$\frac{y_4}{n_4}$
Teve Sinistros	$\frac{y_5}{n_5}$	$\frac{y_6}{n_6}$	$\frac{y_7}{n_7}$	$\frac{y_8}{n_8}$

Quadro 3. 1- Padrões de covariáveis formados no exemplo em que o modelo tem apenas duas variáveis

Tem-se então 8 padrões de covariáveis, de acordo com a sinistralidade e a categoria automóvel.

Os n_i são o número de observações no grupo i, os Y_i a variável aleatória que representa a frequência absoluta de sucessos no grupo i e os $\overline{Y}_i$ a variável aleatória que representa a frequência relativa de sucessos no grupo i.

Se as n_i observações forem independentes em cada grupo e todas elas tiverem a mesma probabilidade π_i de terem o atributo de interesse, então a distribuição de Y_i é binomial com parâmetros π_i e n_i,

$$Y_i \sim B(n_i, \pi_i) \tag{3.7}$$

Logo tem-se para as frequências relativas:

$$\overline{Y}_i \sim B(n_i, \pi_i)/n_i \tag{3.8}$$

A média e a variância para $\overline{Y}_i$ são dadas respectivamente por:

$$E(\overline{Y}_i) = \pi_i \tag{3.9}$$

$$V(\overline{Y}_i) = \frac{V(Y_i)}{n_i^2} = \frac{\pi_i(1-\pi_i)}{n_i} \tag{3.10}$$

Assim, no caso de dados agrupados, as variáveis resposta a considerar são as frequências relativas de sucessos em cada padrão de covariáveis, i.e,

$$\overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \tag{3.11}$$

onde Y_{ij} é uma variável indicadora que toma valor 1 ou 0 se a j-ésima observação no grupo i é um sucesso ou insucesso respectivamente.

Concluindo, pode-se usar a distribuição de Bernoulli para os dados (binários) individuais, ou a distribuição binomial para os dados agrupados, e tomar como resposta a frequência relativa em cada grupo.

As duas abordagens levam exatamente aos mesmos resultados, ou seja, levam à mesma função de verosimilhança e portanto às mesmas estimativas e erros-padrão.

Como todas as variáveis explicativas neste estudo são categóricas então optou-se por agrupar os dados pois desta maneira o conjunto de dados torna-se muito menor.

Voltando agora ao motivo pelo qual o modelo linear gaussiano não se adaptaria bem neste tipo de dados binários, suponha-se que a média era modelada linearmente:

$$\pi(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3.12}$$

Isto conduziria a um desajustamento visto que o termo do lado esquerdo varia entre 0 e 1 e o termo do lado direito da equação toma valores entre $-\infty$ e $+\infty$.

Perante esta situação, é necessário relacionar a probabilidade da resposta $\pi(x)$ com o preditor linear:

$$\varphi = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3.13}$$

via uma função g: $\varphi = g(\mu) = g(\pi)$.

Cada modelo é definido pela escolha de uma função de ligação.

Para decidir qual a função de ligação para o modelo de regressão logística, notemos que a razão entre a probabilidade do sucesso e a probabilidade do insucesso para a resposta varia entre 0 e $+\infty$:

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i}, \tag{3.14}$$

Aplicando a transformação logarítmica (de base e), passamos a ter:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{3.15}$$

que já varia entre $-\infty$ e $+\infty$.

A função de ligação escolhida é então

$$g(\pi) = \text{logit}(\pi), \tag{3.16}$$

Isto é, a equação linear de interesse (linear nos parâmetros de regressão) é:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3.17}$$

Desenvolvendo a equação (3.17) obtém-se:

$$\pi(x) = P\left(Y = 1 \middle| X_1 = x_1, \ldots, X_p = x_p\right)$$

$$= \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} \tag{3.18}$$

3.3 Método de máxima verosimilhança

O método de máxima verosimilhança permite estimar os parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ do modelo de regressão logística.

Tome-se uma amostra aleatória $y = (y_1, y_2, \dots, y_n)$ com função densidade de probabilidade $f(y_i|x_i)$, em que x_i são os valores para as variáveis explicativas em causa.

Ora como a componente aleatória deste modelo é dada por (3.3), então a função de probabilidade de Y condicionada por $X = x_i$ é definida como:

$$f(y_i|x_i) = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \tag{3.19}$$

Assim a função densidade de probabilidade conjunta é dada por:

$$\prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \tag{3.20}$$

Isto é, a função de densidade de probabilidade conjunta é simplesmente o produto das densidades de cada uma das observações e corresponde à função de verosimilhança.

Ora para estimarmos os parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ é necessário maximizar a função de verosimilhança em ordem a cada parâmetro do modelo.

Uma maneira de simplificar este procedimento é tomando o logaritmo da verosimilhança (3.20) [6].

Tem-se que:

$$\log\left(\prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}\right) =$$

$$= \sum_{i=1}^{n} \left(\log(\pi(x_i)^{y_i}) + \log\left(\left(1 - \pi(x_i)\right)^{1-y_i}\right)\right.$$

$$= \sum_{i=1}^{n} \left(y_i \log(\pi(x_i)) + (1 - y_i)\log\left(1 - \pi(x_i)\right)\right.$$

$$= \sum_{i=1}^{n} \left(y_i \log(\pi(x_i)) + \log\left(1 - \pi(x_i)\right) - y_i\log\left(1 - \pi(x_i)\right)\right.$$

$$= \sum_{i=1}^{n} \left(\log\left(1 - \pi(x_i)\right) + y_i \log\left(\frac{\left(\pi(x_i)\right)}{\left(1 - \pi(x_i)\right)}\right)\right) \tag{3.21}$$

Seja $X = (X_0, X_1, \dots, X_p)$ o vector de variáveis explicativas em que X_0 toma valor 1.

Tem-se que:

$$\log\left(\frac{(\pi(x_i))}{(1-\pi(x_i))}\right) = X^t\beta$$

$$\Updownarrow$$

$$\pi(x_i) = \frac{e^{X^t\beta}}{1+e^{X^t\beta}} \tag{3.22}$$

Substituindo (3.22) na expressão (3.21) tem-se então:

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\log\left(1-\frac{e^{X^t\beta}}{1+e^{X^t\beta}}\right)+y_i\log\left(\frac{\frac{e^{X^t\beta}}{1+e^{X^t\beta}}}{\left(1-\frac{e^{X^t\beta}}{1+e^{X^t\beta}}\right)}\right)\right)\\
&\quad=\sum_{i=1}^{n}(-\log(1+e^{X^t\beta})+y_i\log(e^{X^t\beta}))\\
&\quad=\sum_{i=1}^{n}(-\log(1+e^{X^t\beta})+y_iX^t\beta)\\
&\quad=\sum_{i=1}^{n}\sum_{k=0}^{p}y_ix_{k,i}\beta_k-\sum_{i=1}^{n}\log\left(1+e^{\sum_{k=0}^{p}x_{k,i}\beta_k}\right)\\
&\quad=LL(\beta;y,x)
\end{aligned} \tag{3.23}$$

Para podemos encontrar os extremos de LL em $\beta=\hat{\beta}$ é necessário derivar o logaritmo da verosimilhança em função de cada parâmetro e igualar a zero:

$$\frac{\partial LL}{\partial\beta_j}(\hat{\beta}) = 0, \quad j = 0,1,\dots,p \tag{3.24}$$

Desenvolvendo a equação (3.24) tem-se então:

$$\sum_{i=1}^{n}y_ix_{j,i}-\sum_{i=1}^{n}\frac{e^{\sum_{k=0}^{p}x_{k,i}\beta_k}}{1+e^{\sum_{k=0}^{p}x_{k,i}\beta_k}}x_{j,i}=0, \quad j=0,1,\dots,p \tag{3.25}$$

Assim vamos ter p+1 equações que formam um sistema não linear nos parâmetros $\beta_0,\beta_1,\beta_2,\dots,\beta_p$.

Seja,

$$V=\begin{bmatrix}1 & x_{11} & \dots & x_{p1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{1n} & \dots & x_{pn}\end{bmatrix}$$

uma matriz, cuja primeira coluna consta de 1's e cada uma das p colunas seguintes consiste das n observações de cada uma das p variáveis explicativas.

E defina-se $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_n)$, como sendo o vector de probabilidades previstas individuais.

Então o sistema anterior de p+1 equações (equações de verosimilhança) toma a seguinte forma:

$$yV^t = \hat{\pi}V^t \tag{3.26}$$

Pela equação (3.26) tem-se que para j=0:

$$y_1 + y_2 + \cdots + y_n = \widehat{\pi_1} + \widehat{\pi_2} + \cdots + \widehat{\pi_n}$$

$$\Updownarrow$$

$$y_1 - \widehat{\pi_1} + y_2 - \widehat{\pi_2} + \cdots + y_n - \widehat{\pi_n} = 0$$

$$\Updownarrow$$

$$\sum_{i=1}^{n} \left(y_i - \pi(x_i)\right) = 0 \tag{3.27}$$

Pela mesma lógica, para j= 1, 2,…, p, tem-se:

$$\sum_{i=1}^{n} x_{ij}\left(y_i - \pi(x_i)\right) = 0 \tag{3.28}$$

Em particular, resulta da equação (3.27) que para qualquer modelo,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}\,(x_i) \tag{3.29}$$

Ou seja, a soma dos valores observados é igual à soma dos valores previstos pelo modelo [6].

Como as equações de verosimilhança formam um sistema sem solução fechada então é necessária a utilização de métodos iterativos para a sua resolução.

3.3.1 Métodos iterativos para estimação dos parâmetros β

Para resolver as p+1 equações de verosimilhança (3.26), existem vários métodos iterativos. No entanto, vou apresentar somente o Método de Estimação por Máxima Verosimilhança[1] [7].

Dada uma estimativa inicial $\hat{\beta}$ dos parâmetros, podemos calcular o valor do preditor linear $\hat{\varphi} = X_i\hat{\beta}$ e os valores ajustados $\hat{\mu} = \text{logit}^{-1}(\hat{\varphi})$. Com estes valores calculamos de seguida a variável dependente de trabalho[2] z:

[1] Maximun Likelihood Estimation
[2] working dependent variable

$$z = \widehat{\varphi}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} n_i \tag{3.30}$$

onde os n_i são os coeficientes binomiais $\binom{n}{p}$.

De seguida calcula-se a estimativa dos mínimos quadrados ponderados:

$$\hat{\beta} = (V'WV)^{-1}V'Wz \tag{3.31}$$

onde W é a matriz diagonal de pesos com as seguintes entradas:

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i \tag{3.32}$$

A estimativa resultante de β é usada para obter valores ajustados melhorados e o processo é repetido até convergir.

Valores iniciais adequados podem ser obtidos aplicando a função logarítmica aos dados. Para evitar problemas com contagens de 0 ou n_i (o que acontece sempre com dados não agrupados e com valores dicotómicos 0 e 1), calculam-se logits empíricos adicionando $1/2$ a ambos o denominador e numerador:

$$z_i = \log\left(\frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}\right) \tag{3.33}$$

Substituindo (3.33) em (3.31) obtém-se então uma estimativa inicial para $\hat{\beta}$.

A estimativa resultante final é consistente e a sua variância é dada por:

$$\mathrm{var}(\hat{\beta}) = (V'WV)^{-1} \tag{3.34}$$

onde W é a matriz de pesos avaliada na última iteração.

3.4 Interpretação dos parâmetros β do modelo

Segue-se agora a interpretação dos parâmetros β do modelo para diferentes tipos de variáveis explicativas $X = (X_1, X_2, \dots, X_p)$. Estas variáveis podem ser categóricas ou contínuas e considerar-se-á o caso em que o modelo não tem interações.

3.4.1 Caso em que a variável explicativa é contínua

Primeiramente considere-se a situação em que X_i é contínua.

Ao aumentar a coordenada i de $x = (x_1, \dots, x_i, \dots, x_p)$ em c unidades tem-se:

$$x^{(+c)_i} = (x_1, \dots, x_i + c, \dots, x_p) \tag{3.35}$$

Desta maneira, recorrendo à definição de odds ratio e tomando o seu logaritmo obtém-se:

$$\begin{aligned} \log\left(OR\left(x^{(+c)_i}, x\right)\right) &= \log\left(\frac{\text{odds}\left(x^{(+c)_i}\right)}{\text{odds}(x)}\right) \\ &= \log\left(\frac{\frac{\pi\left(x^{(+c)_i}\right)}{1-\pi\left(x^{(+c)_i}\right)}}{\frac{\pi(x)}{1-\pi(x)}}\right) \\ &= \text{logit}\left(x^{(+c)_i}\right) - \text{logit}(x) \\ &= (x_i + c)\beta_i - x_i\beta_i \\ &= c\beta_i \end{aligned} \tag{3.36}$$

Logo,

$$OR\left(x^{(+c)_i}, x\right) = e^{c\beta_i} \tag{3.37}$$

Assim, $e^{c\beta_i}$ corresponde ao odds ratio para o sucesso (Y=1) por aumento de c unidades em X_i, ajustado para as restantes variáveis [6].

3.4.2 Caso em que todas as variáveis explicativas são contínuas

Considerando agora a situação em que $X_1, X_2, \dots, X_p$ são contínuas.
Dadas duas observações, $x_i = (x_{i1}, \dots, x_{ip})$ e $x_j = (x_{j1}, \dots, x_{jp})$ tem-se:

$$\begin{aligned} \log\left(OR\left(x_i, x_j\right)\right) &= \text{logit}(x_i) - \text{logit}\left(x_j\right) \\ &= \left(x_i - x_j\right)\beta \end{aligned} \tag{3.38}$$

E portanto:

$$\begin{aligned} OR\left(x_i, x_j\right) &= e^{\sum_{k=1}^{p}(x_i - x_j)\beta} \\ &= e^{\sum_{k=1}^{p}(x_{ik} - x_{jk})\beta_k} \end{aligned} \tag{3.39}$$

que representa o odds ratio para o sucesso (Y=1), quando se passa das características x_j para x_i.

3.4.3 Caso em que a variável explicativa é dicotómica

Quando X_i é dicotómica, assumindo, sem perda de generalidade que os níveis da variável estão codificados em 0 e 1, tem-se então:

$$\begin{aligned}\log\big(\mathrm{OR}(X_i=1,X_i=0)\big) &= \log\left(\frac{\frac{\pi(X_i=1)}{1-\pi(X_i=1)}}{\frac{\pi(X_i=0)}{1-\pi(X_i=0)}}\right)\\ &= \mathrm{logit}(X_i=1)-\mathrm{logit}(X_i=0)\\ &= \beta_0+\cdots+\beta_{i-1}X_{i-1}+\beta_i+\beta_{i+1}X_{i+1}+\cdots+\beta_pX_p\\ &\quad -\beta_0-\cdots-\beta_{i-1}X_{i-1}-\beta_{i+1}X_{i+1}-\cdots-\beta_pX_p\\ &= \beta_i\end{aligned} \tag{3.40}$$

E portanto,

$$\mathrm{OR}(X_i=1,X_i=0)=e^{\beta_i} \tag{3.41}$$

o que significa que é mais (ou menos) provável e^{β_i} vezes que o sucesso ocorra nos indivíduos com $X_i=1$ do que nos indivíduos com $X_i=0$, mantendo as restantes variáveis explicativas constantes [6].

3.4.4 Caso em que a variável explicativa é politómica

No caso em que X_i é politómica com k categorias, esta é representada por k-1 variáveis dummy, sendo uma das categorias, a categoria de referência (normalmente a primeira, designada por 0).

O que o modelo de regressão logística faz nesta situação é estimar o odds ratio para cada uma das categorias 1, 2,…, k-1, tendo a categoria 0 como grupo de referência.

A constante β_0 corresponde ao log-odds de um indivíduo com zero valores em todas as variáveis explicativas, o que significa que e^{β_0} representa o odds para o sucesso (Y=1) na ausência de variáveis explicativas.

Se o modelo de regressão tiver apenas uma variável explicativa então nesta situação, o OR designa-se por OR bruto. Por outro lado, se o modelo de regressão tiver mais do que uma variável explicativa então o OR designa-se por OR ajustado [6].

3.5 Qualidade do ajustamento

3.5.1 Desviância

A desviância é uma medida baseada no critério da razão de verosimilhanças que permite avaliar a qualidade do ajustamento de um certo modelo tendo em conta o modelo saturado.

A desviância de um modelo linear generalizado com componente aleatória contendo uma distribuição pertencente à família exponencial é:

$$D = 2\sum_{i=1}^{n}(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)) \tag{3.42}$$

Para compreender o que são os parâmetros θ e a função b, seja então Y uma variável aleatória. A distribuição de probabilidade de Y pertence à família exponencial com parâmetro de dispersão se a sua função (densidade) de probabilidade é da forma:

$$f(y|\theta,\varphi) = \exp\left[\frac{1}{a(\varphi)}\left(\sum_{k=1}^{q}\theta_k T_k(y) - b(\theta)\right) + c(y,\varphi)\right] \tag{3.43}$$

com $y \in \mathbb{R}^2$

onde:

- $\theta = (\theta_1, \theta_2, \dots, \theta_q)$ são essencialmente parâmetros de localização, designados por parâmetros canónicos.
- φ é um parâmetro de escala, designado por parâmetro de dispersão.
- As funções $T_1, \dots, T_q$ dependem apenas da variável aleatória Y.
- A função b depende apenas dos parâmetros θ.
- A função a depende apenas do parâmetro de dispersão φ.
- A função c depende apenas da variável aleatória Y e do parâmetro de dispersão φ.

No caso da distribuição binomial $Y \sim B(n,p)$ e portanto:

$$\begin{aligned} f(y,n,p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left[y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right] \end{aligned} \tag{3.44}$$

Assim, para a distribuição binomial tem-se:

$$\begin{aligned} &\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right), b(\theta_i) = -n_i\log(1-\pi_i), \varphi = 1, \\ &T(Y) = Y, c(y,\varphi) = \log\binom{n}{y}, a(\varphi) = 1 \end{aligned} \tag{3.45}$$

Note-se que no modelo saturado tem-se que $\pi_i = \frac{y_i}{n}$ e $\hat{\mu}_i = n\hat{\pi}_i$ ($\hat{\mu}_i$ é o valor ajustado da observação i), e portanto pode-se escrever a desviância na seguinte forma:

$$D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (n - y_i)\log\left(\frac{n - y_i}{n - \hat{\mu}_i}\right)\right] \tag{3.46}$$

A desviância avalia o valor absoluto da diferença entre os valores observados (que coincidem com os valores ajustados para o modelo saturado) e os valores ajustados pelo modelo em estudo.

Como esta estatística compara a log-verosimilhança do modelo de interesse com a log-verosimilhança do modelo saturado então um modelo com ajustamento perfeito tem desviância igual a zero.

No caso em que os dados se encontram agrupados, os valores ajustados são calculados para cada padrão de covariáveis e dependem da probabilidade estimada desse padrão de covariáveis. Nesta situação a estatística desviância converge assimptoticamente (à medida que os tamanhos n_i dos grupos se tornam maiores) para a distribuição $\chi^2(K - (p + 1))$, onde K é o número de grupos e p é o número de parâmetros do modelo excluindo a constante [1].

Já no caso em que os dados são individuais, ou seja, $K = n$ e $n_i = 1$ para todo o $i = 1, \dots, n$, a estatística desviância não converge para nenhuma distribuição conhecida. Além disso, como neste caso a fórmula da desviância pode ser escrita apenas à custa dos valores previstos $\hat{y}_i$'s, essa fórmula não permite comparar valores ajustados com valores observados y_i's pelo que a desviância não pode ser usada para avaliar a qualidade do ajustamento do modelo [1].

3.5.2 Estatística χ^2 de Pearson

Além da desviância existe uma medida alternativa da avaliação da qualidade de ajustamento de um modelo, a estatística χ^2 de Pearson:

$$\chi^2 = \sum_{i=1}^{K}\frac{(O_i - E_i)^2}{E_i} \tag{3.47}$$

onde O_i representa as contagens observadas e E_i as contagens esperadas para a observação i, onde a soma é feita sobre todos os padrões de covariáveis.

Considerando o caso em que a resposta é binomial, então ocorrem duas situações:

- O_i dá o número de sucessos (Y=1) no padrão
- O_i dá o número de insucessos (Y=0) no padrão

Para um número K de padrões de covariáveis tem-se, respetivamente para as duas situações:

- $O_i = y_i$ e $E_i = n\hat{\pi}_i$
- $O_i = n - y_i$ e $E_i = n(1 - \hat{\pi}_i)$

Substituindo em (3.47) obtém-se:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{K} \frac{(y_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} + \frac{(y_i - n\hat{\pi}_i)^2}{n(1-\hat{\pi}_i)} \\ &= \sum_{i=1}^{K} \frac{(y_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i(1-\hat{\pi}_i)} \\ &= \sum_{i=1}^{K} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\pi}_i)} \\ &= \sum_{i=1}^{K} \frac{(y_i - \hat{\mu}_i)^2}{n\hat{\pi}_i(1-\hat{\pi}_i)} \end{aligned} \quad (3.48)$$

Cada termo da soma em (3.48) corresponde ao quadrado da diferença entre valores observados e esperados, dividido pela variância estimada de y_i.

No caso em que os dados são agrupados e as contagens esperadas são todas superiores a 5, a estatística de Pearson segue a distribuição $\chi^2 \sim^a \chi^2(K - (p + 1))$,

onde K é o número de padrões de covariáveis e p é o número de parâmetros do modelo excluindo a constante [1].

Neste caso, χ^2 é assimptoticamente equivalente à desviância.

No entanto para dados individuais, a estatística χ^2 não pode ser usada como teste à qualidade do ajustamento do modelo.

Note-se ainda que a diferença entre estatísticas de Pearson não pode ser usada para comparação de modelos encaixados, contrariamente ao que acontece com a desviância. O motivo pelo qual não pode ser usada deve-se ao facto de a distribuição para a diferença entre as estatísticas de Pearson para modelos encaixados não ser conhecida.

Em geral, a estatística χ^2 toma valores próximos da desviância. No entanto, a existência de uma variável explicativa contínua fará aumentar significativamente o número de padrões de covariáveis e assim comprometer as convergências das estatísticas da desviância e de Pearson e consequentemente, os valores-p destas estatísticas deixam de ser fiáveis [6] e [1].

3.5.3 Teste de Hosmer-Lemeshow

Quando o número de padrões de covariáveis K é próximo do número de observações n, pode-se usar o teste de Hosmer-Lemeshow como alternativa.

Hosmer e Lemeshow propuseram colapsar a tabela de contingência $K\times 2$ (as duas colunas são definidas por a resposta binária y_i e as K linhas os padrões de covariáveis) em uma tabela de contingência $g\times 2$, agrupando para isso as observações de acordo com as probabilidades estimadas $\hat{\pi}_i$ [6]. O agrupamento pode ser feito com base nos valores fixos ou nos percentis das probabilidades estimadas. No primeiro método, escolhe-se pontos de corte k/g , $k = 1, \dots, g-1$, e os grupos contêm todos os indivíduos cujas probabilidades estimadas se encontram entre pontos de corte adjacentes. No segundo método, o primeiro grupo contém os $n'_1 = n/10$ indivíduos que têm as probabilidades estimadas mais baixas e o último grupo contém os $n'_{10} = n/10$ indivíduos que têm as probabilidades estimadas mais altas.

As frequências observadas para as duas colunas da tabela correspondem ao número de respostas positivas ($y_i = 1$) e ao número de respostas negativas ($y_i = 0$) em cada grupo.

As frequências estimadas esperadas são calculadas, somando as probabilidades estimadas de sucesso ($\hat{\pi}_i$) e as de insucesso ($1 - \hat{\pi}_i$) para todos os grupos g. De seguida, a estatística χ^2 de Pearson é calculada a partir da tabela $g\times 2$ de frequências observadas e estimadas.

A estatística de Hosmer-Lemeshow é definida por:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \overline{\pi}_k)^2}{n'_k \overline{\pi}_k (1 - \overline{\pi}_k)} \tag{3.49}$$

onde n'_k é o número total de indivíduos no k-ésimo grupo, c_k o número de padrões de covariáveis no k-ésimo decil,

$$o_k = \sum_{j=1}^{c_k} y_j \tag{3.50}$$

o_k é o número total de respostas dentro do grupo k. E,

$$\overline{\pi}_k = \sum_{j=1}^{c_k} \frac{n_j \hat{\pi}_j}{n_k} \tag{3.51}$$

é a média das probabilidades estimadas no grupo k.

Se o modelo de regressão logística ajustado for o modelo correto, então a distribuição de $\hat{C}$ para os dois tipos de agrupamento aproxima-se à distribuição χ^2 com $g - 2$ graus de liberdade, isto no caso em que as frequências estimadas esperadas ≥ 5. Se

alguma das frequências estimadas esperadas for inferior a 5, então as linhas adjacentes devem ser combinadas até a condição estar satisfeita; o número de graus de liberdade tem de ser reduzido conformemente.
O método de agrupamento baseado nos percentis é preferido. O número de grupos criados recomendados são $g = 10$, no entanto, se as frequências esperadas em alguns dos grupos forem muito pequenas, a estatística de Hosmer-Lemeshow não é fiável pelo que neste caso devemos especificar um número menor de grupos. Contudo, não se pode utilizar menos de 3 grupos uma vez que neste caso é impossível obter a estatística de Hosmer-Lemeshow.

3.6 Testes de significância para os coeficientes do modelo

Depois de os coeficientes de regressão do modelo se encontrarem estimados o próximo passo é avaliar a significância das variáveis explicativas do modelo.
Os testes de hipóteses que se seguem permitem testar essa mesma significância estatística.

3.6.1 Teste da razão de verosimilhanças para modelos encaixados

O teste da razão de verosimilhanças para modelos encaixados consiste em avaliar a significância estatística do conjunto de variáveis usadas no modelo.
Assim, realiza-se sob as seguintes hipóteses:

H_0: $\beta_1 = \cdots = \beta_p = 0$

H_1: pelo menos um dos coeficientes é não nulo

Como o nome do teste indica, esta estatística é baseada na razão de verosimilhanças do modelo nulo e do modelo com as p variáveis explicativas. A estatística de teste é então dada por:

$$G = -2 \ln \left(\frac{\text{verosimilhança para o modelo nulo}}{\text{verosimilhança para o modelo com as p variáveis explicativas}} \right) \tag{3.52}$$

Sob H_0, a estatística G segue uma distribuição χ^2 com p graus de liberdade [6].

3.6.2 Teste de Wald univariado

Rejeitando a hipótese nula do teste da razão de verosimilhanças para modelos encaixados, passa-se então à análise da significância estatística de cada um dos β_j do modelo com j=1, 2, ..,p , individualmente. Ou seja, este teste consiste em testar um submodelo que contém todas as variáveis explicativas do modelo à exceção de uma.
O teste de Wald univariado, realiza-se sob as seguintes hipóteses:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

A hipótese nula indica que o parâmetro de regressão β_j não deve constar do modelo de regressão.
Para amostras de tamanho grande, a estatística de teste é dada por:

$$W_j = \frac{\hat{\beta}_j}{\widehat{se}(\hat{\beta}_j)} \sim^a N(0,1) \tag{3.53}$$

onde, $\widehat{se}(\hat{\beta}_j) = \sigma\sqrt{(X^T X)^{-1}_{jj}}$ é o erro-padrão do coeficiente de regressão β_j. No entanto, σ é desconhecido e por essa razão tem que ser estimado a partir dos dados.
Para valores elevados, em valor absoluto, da estatística W_j pode-se rejeitar a hipótese nula e assim concluir que a variável explicativa X_j é significativa para o modelo em causa [6].

3.6.3 Teste de Score

O teste de Score, realiza-se sob as seguintes hipóteses:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

E a estatística de teste é definida como [6]:

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{3.54}$$

Uma investigação mais profunda, indicou que o teste de Wald é menos eficaz do que o teste de razão de verosimilhanças para modelos encaixados e pode até mesmo apresentar um comportamento aberrante: pode falhar na rejeição da hipótese nula

quando o coeficiente é significante. Por esta razão o teste de razão de verosimilhanças para modelos encaixados é normalmente recomendado [9].
Por outro lado, o teste de Score não requer o ajustamento do modelo com todas as p variáveis explicativas, o que reduz bastante o esforço computacional. Esta redução no esforço computacional é frequentemente citada como a sua maior vantagem [6].

3.7 Estimação de intervalos de confiança

3.7.1 Intervalo de confiança para os coeficientes β_j do modelo

O intervalo mais usual com $100(1-\alpha)\%$ de confiança para um qualquer coeficiente β_j, j= 1, 2, …, p usa a estatística de Wald e portanto corresponde a:

$$\mathrm{IC}_{100(1-\alpha)\%} \text{ para } \beta_j$$

$$= \left(\hat{\beta}_j - \mathrm{N}_{1-\frac{\alpha}{2}}\mathrm{se}(\hat{\beta}_j), \hat{\beta}_j + \mathrm{N}_{1-\frac{\alpha}{2}}\mathrm{se}(\hat{\beta}_j)\right) \tag{3.55}$$

onde, $\mathrm{N}_{1-\frac{\alpha}{2}}$ representa o $\left(1-\frac{\alpha}{2}\right)$ quantil da distribuição normal reduzida, $\mathrm{N}(0,1)$ [10] e [6].
Suponhamos agora que existem dados dispersos, em que os erros-padrão dos coeficientes são sobre-estimados e portanto os valores de W_j são muito baixos, conduzindo a falsas falhas de significância estatística de W_j, também conhecido como efeito de Hauck-Donner [11]. Ora nesta situação não é fiável usar o intervalo de confiança anteriormente descrito (3.55) e por este motivo o intervalo de confiança preferido é o intervalo de confiança de perfil baseado na verosimilhança.

3.7.2 Intervalo de confiança de perfil baseado na verosimilhança

Para um qualquer coeficiente β_j, j= 1, 2, …, p, podemos definir um perfil de máxima verosimilhança, $\mathrm{l}_p(\beta_j)$, maximizando para isso o logaritmo de máxima verosimilhança sobre os restantes parâmetros enquanto que β_j é fixado num determinado valor.
O intervalo de confiança de perfil baseado na verosimilhança para β_j é obtido considerando valores de β_j que são razoavelmente consistentes com os dados. Mais

especificamente, o intervalo com 95% de confiança anteriormente referido é dado pelo conjunto de todos os valores β_j que satisfazem:

$$2\times\{l_p(\hat{\beta}_j) - l_p(\beta_j)\} \leq 3.84 \tag{3.56}$$

onde o valor limite do lado direito da equação é obtido a partir da distribuição χ^2 com 1 grau de liberdade [12].

3.7.3 Intervalo de confiança para o Odds Ratio

Agora se quisermos calcular o intervalo com $100(1-\alpha)\%$ de confiança para o Odds Ratio, observe-se que:

$$OR\left(x^{(+1)_i}, x\right) = e^{\beta_i} \tag{3.57}$$

Logo, basta tomar a exponencial de cada um dos extremos do intervalo de confiança para os coeficientes que usa a estatística de Wald ou de cada um dos extremos do intervalo de confiança de perfil baseado na verosimilhança [13].
Assim tem-se:

$$IC_{100(1-\alpha)\%} \text{ para } OR\left(x^{(+1)_i}, x\right)$$
$$= \left(e^{\hat{\beta}_j - N_{1-\frac{\alpha}{2}} se(\hat{\beta}_j)}, e^{\hat{\beta}_j + N_{1-\frac{\alpha}{2}} se(\hat{\beta}_j)}\right) \tag{3.58}$$

A estimativa obtida para o OR será estatisticamente significativa sempre que o correspondente intervalo de confiança não contiver o valor 1.
Note-se que enquanto que os intervalos de confiança para os coeficientes são centrados na estimativa do coeficiente, os intervalos de confiança para o odds ratio não são centrados no odds ratio estimado, por causa da função exponencial.

3.7.4 Intervalo de confiança para a probabilidade de sucesso da resposta

Os intervalos de confiança para o logit usam a matriz de co-variância do estimador $\hat{\beta}$.
A $\widehat{var}\left(\hat{\pi}_i(x)\right) = x^t \widehat{var}\left(\hat{\beta}\right) x$, e por isso obtém-se que o intervalo de predição para $\pi(x_i)$ (probabilidade de sucesso da resposta) com $100(1-\alpha)\%$ de confiança é [6]:

$$\text{IC}_{100(1-\alpha)\%} \text{ para } \pi(x_i)$$

$$= \left(\hat{\pi}_i(x) - N_{1-\frac{\alpha}{2}} \text{se}\left(\hat{\pi}_i(x)\right), \hat{\pi}_i(x) + N_{1-\frac{\alpha}{2}} \text{se}\left(\hat{\pi}_i(x)\right) \right) \quad (3.59)$$

3.8 Curva ROC

Para avaliar a precisão das previsões do modelo de regressão logística, uma ferramenta poderosa é a análise da curva ROC ("Receiver Operating Characteristic"). Esta análise é feita por meio de um método gráfico simples e robusto que permite estudar a variação da sensibilidade e especificidade para diferentes valores de corte.
Como a resposta é binária (1 se o cliente anula o contrato de seguro e 0 caso contrário) então é empregada uma regra de decisão baseada em escolher um ponto de corte de forma que um indivíduo com valores previstos maiores que o ponto de corte é classificado como tendo anulado o contrato e um indivíduo com valores previstos menores que o ponto de corte é classificado como não tendo anulado o contrato [14].
Após o ajuste de um modelo e a determinação do ponto de corte, é importante avaliar o poder de discriminação do modelo, isto é, discriminar os eventos dos não eventos.
Para melhor compreensão, considere-se a seguinte tabela em que se encontra resumida a informação sobre os valores observados e previstos pelo modelo [15]:

Valor Previsto	**Valor Observado**		
	Y=1	**Y=0**	**Total**
Y=1	Verdadeiros Positivos (VP)	Falsos Positivos (FP)	VP+FP
Y=0	Falsos Negativos (FN)	Verdadeiros Negativos (VN)	FN+VN
Total	VP+FN	FP+VN	VP+FP+FN+VN

Quadro 3. 2 - Valores observados vs valores previstos pelo modelo

Um Verdadeiro Positivo (VP) neste caso de estudo é a probabilidade de um indivíduo que anulou o contrato de seguro ser classificado como tendo realmente anulado o contrato.

Por outro lado, um Verdadeiro Negativo (VN) é a probabilidade de um indivíduo que não anulou o contrato de seguro ser classificado como não tendo anulado o contrato.

Já um Falso Positivo (FP) é a probabilidade de um indivíduo que não anulou o contrato ser classificado como tendo anulado o contrato.

Por fim um Falso Negativo (FN) é a probabilidade de um indivíduo que anulou o contrato ser classificado como não tendo anulado o contrato.

Temos então que a sensibilidade é definida como a probabilidade do modelo fornecer um resultado positivo, dado que o indivíduo realmente anulou o contrato:

$$\mathrm{S_E} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \tag{3.60}$$

Por outro lado, a especificidade é definida como a probabilidade do modelo fornecer um resultado negativo, dado que o indivíduo não anulou o contrato:

$$\mathrm{E_S} = \frac{\mathrm{VN}}{\mathrm{FP} + \mathrm{VN}} \tag{3.61}$$

Pelas fórmulas da sensibilidade e da especificidade pode-se concluir que as duas medidas são independentes entre si. No cálculo da sensibilidade utiliza-se apenas os indivíduos que anularam o contrato e no cálculo da especificidade utiliza-se somente os indivíduos que não anularam o contrato.

No gráfico seguinte pode-se ver um exemplo de como variam os valores da sensibilidade e da especificidade para as curvas ROC de diferentes classificadores assim como as respetivas distribuições para a variável resposta e diferentes pontos de corte:

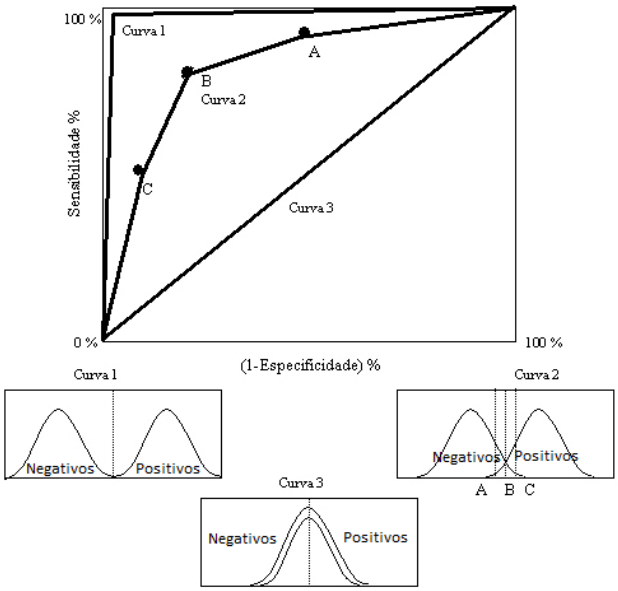


Figura 3. 1 - Curvas ROC para diferentes tipos de classificadores e respetivas distribuições

A curva 3 é um mau classificador e corresponde a um modelo que dá resultados positivos ou negativos aleatoriamente, ou seja, não tem utilidade prática [13].

Um classificador considerado bom é o da curva 2.

O classificador correspondente à curva 1 dificilmente é alcançado e representa o modelo que classifica corretamente todos os casos positivos e negativos, ou seja, o modelo perfeito.

Relativamente ao classificador considerado bom (curva2), note-se que quanto maior é o ponto de corte (ponto C) maior é a especificidade e menor a sensibilidade. Por outro lado, quanto menor for o ponto de corte (ponto A) maior é a sensibilidade e menor a especificidade.

A área sob a curva ROC (AUC) é uma medida de discriminação e serve de comparação entre possíveis modelos [6]:

Valor do AUC	Discriminação do Modelo
0.5	Sem discriminação
[0.7;0.8[	Discriminação aceitável
[0.8;0.9[	Discriminação excelente
$\geq$ **0.9**	Discriminação excepcional

Quadro 3. 3 - Discriminação do modelo segundo o valor do AUC

3.9 AIC e BIC

O critério de informação de Akaike (AIC) e o critério de informação Bayesiana (BIC) são medidas para apoiar a escolha de um modelo entre um conjunto finito de modelos. Estes dois critérios são muito parecidos e próximos entre si.

O AIC é definido como [14] e [3]:

$$\mathrm{AIC} = -2\mathrm{LL}(\beta) + 2\mathrm{p}, \tag{3.62}$$

onde $\mathrm{LL}(\beta)$ é o logaritmo da função de verosimilhança e p é o número de parâmetros do modelo estimado.

O BIC foi desenvolvido por Gideon Schwarz em 1978 e é definido como:

$$\mathrm{BIC} = -2\mathrm{LL}(\beta) + 2\mathrm{p}\log(\mathrm{n}), \tag{3.63}$$

onde n é o número total de observações.

A situação ideal seria selecionar um modelo parcimonioso, ou seja, que estivesse bem ajustado aos dados e que tivesse um número reduzido de parâmetros.

Apesar de nenhum dos dois critérios avaliar o ajustamento do modelo aos dados, estes lidam com o número de parâmetros do modelo. E como ambos os critérios crescem com o aumento do número de parâmetros do modelo então o modelo com menor valor para o critério de informação de Akaike ou com menor valor para o critério de informação Bayesiana será o melhor.

É de notar no entanto que o critério BIC penaliza mais o número de parâmetros do que o critério de AIC pois o número de parâmetros é multiplicado por $\log(\mathrm{n})$ que por sua vez, cresce à medida que o número de observações aumenta.

3.10 Resíduos

Os resíduos exprimem a discrepância entre o valor observado $\mathrm{y_i}$ e o valor $\hat{\mu}_\mathrm{i}$ ajustado pelo modelo [16]. A análise dos resíduos é essencialmente gráfica e permite avaliar a qualidade de ajustamento do modelo no que diz respeito à escolha da função de ligação, à escala usada nas variáveis explicativas e também permite identificar observações influentes e “outliers”.

A seguir são descritos dois tipos de resíduos, os resíduos de Pearson estandardizados e os resíduos da desviância estandardizados.

3.10.1 Resíduos de Pearson estandardizados

Os resíduos de Pearson estandardizados para uma dada observação i são definidos por [7]:

$$PR_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}} \tag{3.64}$$

Estes resíduos calculam a diferença entre os valores observados y_i e os valores ajustados $\hat{\mu}_i$, sendo essa diferença dividida pela estimativa do desvio padrão do valor observado. O denominador resulta do fato da $\mathrm{var}(y_i) = n_i\pi_i(1-\pi_i)$.

Existem agora duas situações:

1) Os dados são agrupados e as dimensões dos padrões de covariáveis são razoáveis
2) Os dados são agrupados e o número de padrões de covariáveis é reduzido

Na primeira situação, um resíduo de Pearson estandardizado com valor superior a 3 sugere um mau ajustamento.

Na segunda situação, um resíduo de Pearson estandardizado com valor superior a 2 sugere um mau ajustamento.

Em ambas as situações, os resíduos seguem aproximadamente uma distribuição normal reduzida.

Já no caso em que os dados são individuais, estes resíduos não seguem aproximadamente uma distribuição normal reduzida. Neste caso em que os dados são individuais pode-se usar como alternativa os resíduos da desviância estandardizados.

3.10.2 Resíduos da desviância estandardizados

Os resíduos mais utilizados quando se trabalha com modelos de regressão logística são os resíduos da desviância[3].

Este tipo de resíduo é baseado na função desviância. Como referido anteriormente, a desviância é usada como uma medida de discrepância nos modelos lineares generalizados. Cada observação i contribui com uma quantidade d_i para esta medida de discrepância e portanto, [17]

$$D(y, \hat{\mu}) = \sum_i d_i \tag{3.65}$$

Os resíduos da desviância são então dados por:

[3] deviance residual

$$R_i^D = \text{sinal}(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{3.66}$$

Como os resíduos apresentam variabilidades diferentes, devem ser estandardizados pelo erro padrão de y_i, considerando também os efeitos das covariáveis nos resíduos através da matriz de projecção (hat).

Assim, os resíduos da desviância estandardizados são definidos por:

$$R_i^{D'} = \frac{R_i^D}{\sqrt{(1 - h_{ii})}} \tag{3.67}$$

em que h_{ii} representa os valores da diagonal da matriz de projecção, que no caso do modelo logístico é definida por:

$$H = V^{1/2}X\left(X^T V X\right)^{-1} X^T V^{1/2}\,, \tag{3.68}$$

onde X representa a matriz $n\times(p+1)$ que contém os valores das p covariáveis e V a matriz diagonal $n\times n$ com entradas dadas por $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$ [18] e [19].

3.11 "Outliers"

Um "outlier" é um indivíduo com observações que não se ajustam ao modelo encontrado, apresentando portanto resíduos grandes quando comparados com os resíduos das restantes observações [20].

Estes pontos devem ser inicialmente identificados e depois detalhadamente examinados, de forma a garantir que as conclusões do modelo não dependem fortemente da presença de observações extremas.

Observações com resíduos estandardizados superiores a 3.3 devem ser analisadas pormenorizadamente, e eventualmente excluídas do modelo. É importante também olhar para o padrão de dispersão gráfica.

3.12 Deteção de observações influentes

Uma observação é influente se a sua exclusão do modelo de regressão provoca uma mudança substancial nos valores ajustados. A deteção destas observações influentes é realizada essencialmente através do cálculo de "Leverages" e Distâncias de Cook.

3.12.1 "Leverages"

Pontos com "leverages"[4] altas são "outliers" da variável explicativa X na medida em que estão muito distanciados de todos os restantes valores de X, contudo estes pontos não correspondem necessariamente a pontos com muita influência sobre a estimação dos parâmetros do modelo. É preciso investigar esses pontos.

Os valores de "Leverage" são dados pelos elementos da diagonal da matriz de projeção, h_{ii}, sendo que a soma destes valores corresponde ao número de parâmetros do modelo, isto é:

$$\sum_{i=1}^{n} h_{ii} = p + 1, \text{ com } h_{ii} \in [0,1] \tag{3.69}$$

Assim, uma observação é considerada influente se $h_{ii} > \frac{2p}{n}$ [21]. À partida, de entre essas observações, aquelas que não apresentarem resíduos grandes não serão preocupantes.

Concluindo, os pontos com "leverages" altas devem ser identificados (não imediatamente excluídos). Graficamente, é usual representar os valores de "Leverage" de cada observação i contra i como representado na figura 3.2:

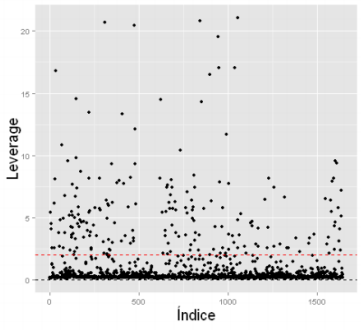


Figura 3. 2 - Gráfico de valores de "Leverage" de cada observação i contra a ordem de observações

3.12.2 Distâncias de Cook

A distância de Cook é usada para avaliar a influência da observação i, ou seja, quantifica o peso dessa observação no modelo.

A distância de Cook para a observação i é definida por:

$$C_i^2 = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(-i)}\right)^2}{(p+1)\overline{\sigma}^2} \tag{3.70}$$

[4] Em português, traduz-se para repercussão ou alavanca

onde $\overline{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p+1)}$ é um estimador não enviesado da variância de Y. E $\hat{y}_{j(-i)}$ representa o valor ajustado para y_j quando o modelo é estimado excluindo a observação i.

Assim, consideram-se observações influentes as observações cuja distância de Cook é superior a $\frac{4}{n}$, onde n é o número de observações [22]. Estes pontos devem ser analisados e o modelo deve ser ajustado sem estes pontos de forma a comparar as estimativas dos parâmetros obtidas. Graficamente, é usual representar os valores de Distância de Cook cada observação i contra i como representado na figura 3.3:

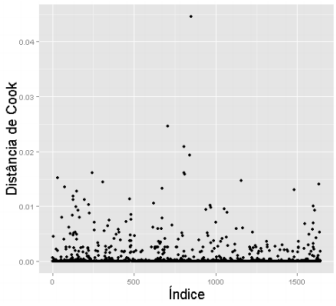


Figura 3. 3 - Gráfico de valores de Distância de Cook de cada observação i contra a ordem de observações

Capítulo 4

Resultados

4.1 Associações entre as variáveis

Antes de incluir variáveis no modelo é importante verificar o nível de associação entre pares dessas variáveis explicativas.

Para isso usou-se a estatística Cramer's V que é uma medida de associação entre duas variáveis categóricas. Esta medida varia entre 0 (correspondente à associação nula entre as variáveis) e 1 (associação total entre as variáveis).

Para o cálculo da estatística Cramer's V são necessários os seguintes dados:

- 1 População
- 2 Fatores (A e B)
- R níveis distintos do fator A
- C níveis distintos do fator B
- $O_{i,j}$ = # indivíduos observados com factor A no nível i e fator B no nível j
- $E_{i,j}$ = # indivíduos esperados com factor A no nível i e factor B no nível j

A medida qui-quadrado é definida como:

$$U = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{4.1}$$

Assim o Cramer's V calcula-se da seguinte forma [23]:

$$V = \sqrt{\frac{U/n}{\min(C-1, R-1)}} \tag{4.2}$$

É de notar que uma associação alta indica apenas associação estatística e não necessariamente causalidade.

Quando existe associação elevada entre as variáveis explicativas isso é refletido, regra geral, em elevados erros padrão dos coeficientes estimados e por esse motivo é importante analisar essas variáveis e se for necessário retirá-las do estudo.

Na tabela seguinte encontra-se a descrição dos níveis de associação para os diferentes valores da estatística Cramer's V [24]:

Valor de Cramer's V	Nível de associação
0	Nula
]0, 0.15]	Muito fraca
]0.15, 0.20]	Fraca
]0.20, 0.25]	Moderada
]0.25, 0.30]	Moderadamente forte
]0.30, 0.35]	Forte
]0.35, 0.40]	Muito forte
]0.40, 0.50]	Super forte
]0.50, 0.99]	Redundante
1	Associação total

Quadro 4. 1 - Níveis de associação para diferentes valores da estatística Cramer's V

Para ver as associações de pares de variáveis para um valor absoluto maior ou igual a 0 da estatística Cramer's V, segue-se o gráfico produzido no programa Emblem:

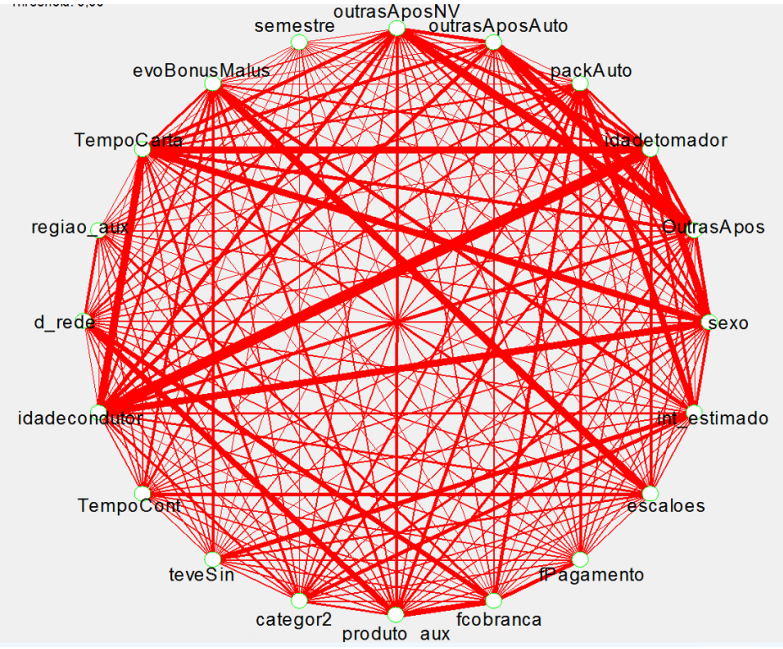


Figura 4. 1 - Associações de pares de variáveis para um valor absoluto ≥0

Note-se que quanto maior é a espessura da linha maior é a associação entre as variáveis.

Na figura 4.1 vemos que as variáveis são todas associadas umas com as outras, no entanto não há problema pois essas associações são na sua grande maioria baixas, ou seja, pouco significativas.

Para visualizar melhor as associações mais significativas segue-se o seguinte gráfico:

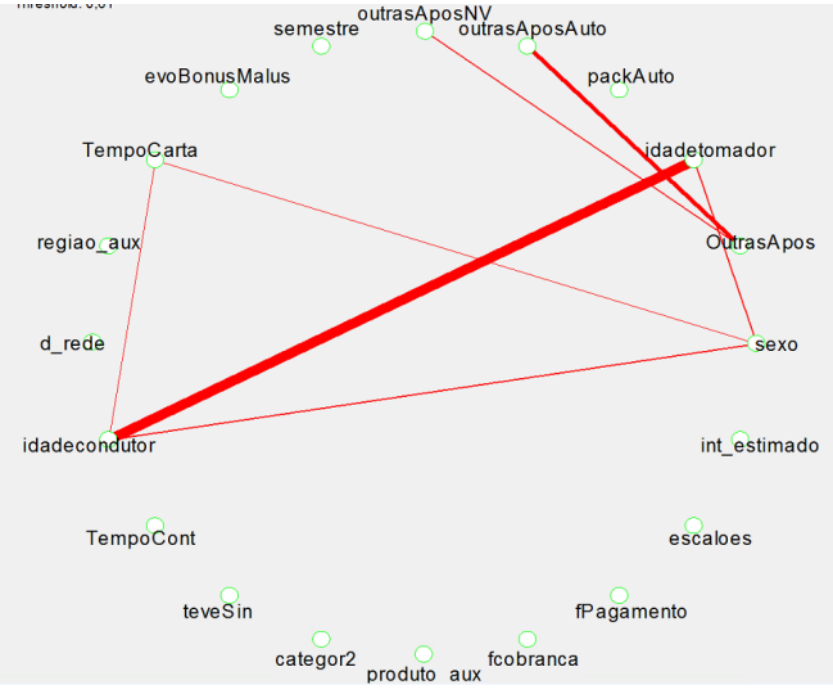


Figura 4. 2 - Associações de pares de variáveis para um valor absoluto ≥0.5

Pela figura 4.2 conclui-se que as associações mais significativas são entre:

1) Idade do Condutor e Idade do Tomador
2) Idade do Condutor e Tempo de Carta
3) Outras Apólices e Outras Apólices não vida
4) Outras Apólices e Outras Apólices Auto
5) Sexo e Idade do Condutor
6) Sexo e Idade do Tomador
7) Sexo e Tempo de Carta

A forte associação entre Idade do Condutor e Idade do Tomador é explicada pelo facto de a maioria dos tomadores do seguro automóvel serem também condutores do veículo seguro. Assim, ter as duas variáveis presentes não traz grande vantagem para o modelo. De entre as duas variáveis, a mais significativa para a seguradora é a Idade do Condutor pelo que não se considerará no modelo a variável Idade do Tomador.

A associação entre as variáveis Idade do Condutor e Tempo de Carta deve-se ao facto de que quanto mais velho é o condutor maior se espera que seja o seu tempo de carta e vice-versa.

A variável Outras Apólices é mais associada com a variável Outras Apólices Automóvel do que com a variável Outras Apólices não Vida. Isto porque a quantidade de apólices automóvel é superior à quantidade de apólices não vida na seguradora. Uma vez que a variável Outras Apólices engloba as outras duas, opta-se então por considerar apenas esta variável no modelo.

A associação entre a variável sexo e as variáveis Idade do Condutor e Tempo de Carta é relativamente significativa pois estas últimas duas têm valor NA se o cliente for empresa e a variável sexo toma valor " jurídico" nessa mesma situação.

Em suma, eliminou-se as variáveis redundantes: Idade do Tomador, Outras Apólices Não Vida e Outras Apólices Automóvel em consequência desta análise.

4.2 Seleção das variáveis a incluir no modelo

4.2.1 Agrupamento de categorias com frequências próximas de zero em algumas das variáveis

A deteção de frequências próximas de zero nas categorias das diferentes variáveis deve ser feita antes do ajuste do modelo de regressão logística, uma vez que esta situação faz com que surjam coeficientes e/ou erros padrão demasiado elevados [13]. A solução passa assim por agregar algumas categorias dos fatores em causa. Por este motivo:

- As categorias redução $\in$]25€;50€] e redução $\in$]20€;25€] da variável Intervalo estimado de variação do prémio foram agrupadas numa única categoria: redução $\in$]20€;50€];
- As categorias Protocolos-AssComerciais, Protocolos-Sonae e Protocolos-Outros da variável Produto foram agrupadas numa única categoria: Protocolos-Outros;
- As categorias até aos 20 anos e dos 21 aos 25 anos da variável Idade do Condutor foram agrupadas numa única categoria: até aos 25 anos.

4.2.2 Divisão dos dados em treino e teste

Para avaliar a capacidade de previsão do modelo de regressão logística, é importante que os dados que são utilizados para ajustar o modelo sejam diferentes dos que são utilizados na avaliação do mesmo.

Assim, dividiu-se os dados num conjunto de treino e num conjunto de teste. O modelo é então ajustado no conjunto de treino e posteriormente aplicado no conjunto de teste para se poder avaliar a confiança que podemos ter na classificação de observações futuras.

Para este estudo, decidiu-se utilizar o procedimento standard e dividir aleatoriamente os dados num conjunto de treino com 70% dos dados e num conjunto de teste com os restantes 30% dos dados, tendo sido este procedimento realizado apenas uma vez.

4.2.3 Classes de referência das variáveis explicativas

As classes de referência escolhidas para cada variável foram aquelas com maior frequência:

Variável	Classe de Referência
Sinistralidade	Não teve sinistros
Antiguidade do Contrato	≤ 1 ano
Escalão de Bónus	7
Intervalo Estimado de Variação do Prémio	0 €
Semestre do Vencimento	1º Semestre
Evolução Bónus-Malus	Sem Evolução
Outras Apólices	Sim
Região	Beira Litoral
Categoria Automóvel	Ligeiros
Pack Automóvel	Responsabilidade Civil
Produto	Protect
Forma de Pagamento	Anual
Forma de Cobrança	Agente cobrador
Rede	RNA
Sexo	Masculino
Idade do Condutor	Mais de 40 anos
Tempo de Carta	Mais de 10 anos

Quadro 4. 2 - Classes de referência das variáveis explicativas

4.2.4 Variáveis Dummy

Como muitas das variáveis explicativas têm mais do que duas categorias distintas então é necessário utilizar variáveis auxiliares designadas por variáveis Dummy ou variáveis indicatrizes. Se uma variável categórica tem k classes então são necessárias k-1 indicatrizes.

O software SAS gera estas variáveis indicatrizes automaticamente, não havendo portanto necessidade de as criar previamente.

4.2.5 Seleção de variáveis e escolha do modelo final

O processo de seleção de variáveis e de escolha do modelo final passou pela comparação de diferentes modelos recorrendo às medidas de AUC e AIC.

Na tabela seguinte encontram-se os diferentes modelos a comparar. As variáveis com "✔" encontram-se incluídas no modelo, enquanto que as variáveis com "⊘" se encontram excluídas do modelo.

Variável / Modelo	Sinistralidade	Antiguidade Do Contrato	Escalão de Bónus	Int. Est. Var. Prémio	Semestre Vencimento	Evolução Bónus-Malus	Outras Apólices	Região	Categoria Automóvel	Pack Automóvel	Produto	Forma de Pagamento	Forma de Cobrança	Rede	Sexo	Idade do Condutor	Tempo de Carta
1	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
2	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
3	⊘	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
4	✔	⊘	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
5	✔	✔	⊘	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
6	✔	✔	✔	⊘	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
7	✔	✔	✔	✔	⊘	⊘	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
8	✔	✔	✔	✔	⊘	✔	⊘	✔	✔	⊘	✔	✔	✔	✔	✔	✔	✔
9	✔	✔	✔	✔	⊘	✔	✔	⊘	✔	⊘	✔	✔	✔	✔	✔	✔	✔
10	✔	✔	✔	✔	⊘	✔	✔	✔	⊘	⊘	✔	✔	✔	✔	✔	✔	✔
11	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	⊘	✔	✔	✔	✔	✔	✔
12	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	⊘	✔	✔	✔	✔	✔
13	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	⊘	✔	✔	✔	✔
14	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	⊘	✔	✔	✔
15	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔
16	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	⊘	✔
17	✔	✔	✔	✔	⊘	✔	✔	✔	✔	⊘	✔	✔	✔	✔	✔	✔	⊘

Quadro 4. 3 - Modelos candidatos a final

A seguir seguem-se os valores de AIC e AUC de cada modelo presente na tabela anterior:

Modelo	AIC	AUC
1	1661663.7	0.7384
2	1661824.3	0.7384
3	1663181.6	0.7378
4	1662972.3	0.7378
5	1665123.1	0.7367
6	1671435.3	0.7334
7	1662103.5	0.7382
8	1799102.0	0.6384
9	1662563.2	0.7380
10	1663059.6	0.7378
11	1664231.0	0.7368
12	1671325.9	0.7335
13	1664507.1	0.7370
14	1662690.1	0.7379
15	1664990.8	0.7367
16	1662942.5	0.7378
17	1662295.3	0.7381

Quadro 4. 4 - Valores das medidas de AIC e AUC para os modelos candidatos a final

É de notar que o modelo a ser escolhido terá que ser razoavelmente simples e terá que ter um bom ajustamento aos dados. Isto implica encontrar um equilíbrio entre melhorar o ajustamento sem necessariamente aumentar a complexidade do modelo.
Tendo isto em conta, observe-se que do modelo 1 para o modelo 2, o AIC aumenta apenas 160.6 e o AUC mantém-se. Assim a contribuição das variáveis Semestre do Vencimento e Pack Automóvel não é importante para o modelo de regressão logística. Esta conclusão já era de prever quando se fez a análise da relação da resposta com cada uma destas variáveis.
Assim, pode-se excluir o modelo 1 da lista de modelos candidatos a final e toma-se agora como referência o modelo 2 na comparação com os restantes modelos propostos.

Comparação do modelo 2 com os restantes modelos			
Modelo	**Aumento no AIC relativamente ao modelo 2**	**Diminuição no AUC relativamente ao modelo 2**	**Decisão**
3	1357.3	6e-04	Manter a variável Sinistralidade no modelo
4	1148	6e-04	Manter a variável Antiguidade do Contrato no modelo
5	3298.8	0.0017	Manter a variável Escalão de Bónus no modelo
6	9611	0.005	Manter a variável Intervalo estimado de variação do prémio no modelo
7	279.2	2e-04	Manter a variável Evolução Bónus-Malus no modelo
8	137277.7	0.1	Manter a variável Outras Apólices no modelo
9	738.9	4e-04	Manter a variável Região no modelo
10	1235.3	6e-04	Manter a variável Categoria Automóvel no modelo
11	2406.7	0.0016	Manter a variável Produto no modelo
12	9501.6	0.0049	Manter a variável Forma de Pagamento no modelo
13	2682.8	0.0014	Manter a variável Forma de Cobrança no modelo
14	865.8	5e-04	Manter a variável Rede no modelo
15	3166.5	0.0017	Manter a variável Sexo no modelo
16	1118.2	6e-04	Manter a variável Idade do Condutor no modelo
17	471	3e-04	Manter a variável Tempo de Carta no modelo

Quadro 4. 5 - Comparação do modelo 2 com os restantes modelos

Resumindo, todas as variáveis exceto Semestre do Vencimento e Pack Automóvel devem ser mantidas no modelo de regressão logística pois a ausência das variáveis explicativas contidas na tabela anterior fazem aumentar significativamente o AIC e levam a uma diminuição no AUC. Além disso estas variáveis são importantes para a realidade da empresa.

Tem-se então que o modelo final é o 2.

4.2.6 Agrupamento de categorias não significativas em algumas variáveis no modelo ajustado

Note-se aqui que sempre que uma variável categórica é introduzida ou excluída de um modelo, todas as suas categorias devem ser consideradas ou excluídas do modelo, respetivamente. Isto significa que, se na tabela dos coeficientes estimados uma das categorias não for significativa, essa categoria não deve ser excluída. A exclusão de uma categoria implicaria uma nova definição da variável categórica. Assim as categorias que não se revelam significativas no modelo ajustado devem ser agregadas com outras categorias.

Desta maneira, a categoria Pontos da variável Rede que não se revelou significativa foi agrupada com a categoria Outros, formando uma única categoria, Outros.

Também a categoria redução $\in$]10€;15€] da variável Intervalo Estimado de Variação do Prémio não se revelou significativa pelo que se agrupou esta com a categoria redução $\in$]5€;10€], dando origem a uma nova categoria, redução $\in$]5€;15€].

Como consequência deste processo de agregação de categorias, o modelo final apresenta agora um valor de AIC=1662042.5 e um valor de AUC=0.7383, sendo portanto um modelo com uma discriminização aceitável.

4.2.7 Inclusão de possíveis interações entre as variáveis

Uma interação entre duas variáveis explicativas ocorre quando a associação entre uma das variáveis explicativas e a resposta é diferente conforme os valores de outra variável explicativa [6].

Como o modelo tem 15 variáveis explicativas então ter-se-ia de testar 105 interações, o que computacionalmente iria exigir muito tempo. Para além disso iria aumentar consideravelmente o número de testes de hipóteses a realizar, o que traria problemas quanto ao nível de significância. Assim, foram escolhidas apenas duas variáveis (tempo de carta e idade do condutor) que na opinião do meu co-orientador pudessem apresentar interação.

Incluindo a interação entre tempo de carta e idade do condutor no modelo, os valores de AIC e de AUC obtidos foram 1661999.2 e 0.7383 respetivamente. Ou seja, o AUC mantém-se e o AIC só diminui 43.3 o que revela que esta interação não traz uma melhoria significativa para o modelo e que por isso não deve ser incluída neste.

4.3 Testes de hipóteses para avaliação da significância estatística dos coeficientes de regressão

Neste ponto são apresentados os resultados para os testes de hipóteses referidos no Capítulo 3 nas secções 3.6.1, 3.6.2 e 3.6.3.

Para o teste da razão de verosimilhanças para modelos encaixados o valor p obtido foi < 0.0001 pelo que com um nível de significância de 0.0001 deve-se rejeitar H_0 e concluir que pelo menos um dos coeficientes é não nulo.

Como a hipótese nula do teste anterior foi rejeitada então passa-se à análise da significância estatística de cada um dos β_j do modelo com $j = 1, 2, .., p$, individualmente, utilizando para isso, o teste de Wald univariado ou o teste de Score.

Para o teste de Wald univariado e para o teste de Score obtiveram-se para cada um dos β_j do modelo com $j = 1, 2, .., p$ um valor $p < 0.0001$ pelo que que com um nível de significância de 0.0001 deve-se rejeitar H_0 e concluir que todas as variáveis são significativas.

4.4 Capacidade de previsão do modelo

A próxima etapa passa por escolher um ponto de corte na curva ROC.

Assim, analisando os possíveis pontos de corte para este modelo, temos as seguintes medidas de sensibilidade e especificidade:

Ponto de corte	Sensibilidade	Especificidade
0.01	100%	0%
0.02	99.8%	1.5%
0.03	99.1%	6.2%
0.04	97.1%	15.5%
0.05	94.4%	26%
0.06	91.5%	34.2%
0.07	88.8%	40.6%
0.08	86.4%	45.3%
0.09	84.1%	49.1%
0.1	82.1%	52.2%
0.2	56.5%	76.5%
0.3	24.2%	93.3%

Quadro 4. 6 - Valores de sensibilidade e especificidade para diferentes valores de corte

Ponto de corte	Sensibilidade	Especificidade
0.4	7.2%	98.6%
0.5	2.2%	99.7%
0.6	0.6%	99.9%
0.7	0.2%	100%
0.8	0	100%
0.9	0	100%

Quadro 4. 6 – Continuação da página anterior

Tendo analisado cuidadosamente o quadro 4.6, o ponto de corte escolhido foi o 0.06 pois para a seguradora é preferível que o modelo consiga acertar no maior número possível de observações de clientes que anulam o contrato de seguro, mantendo ainda assim, uma probabilidade razoável de acertar nas observações em que o cliente não anula o contrato de seguro. Assim, para o ponto de corte 0.06 tem-se uma sensibilidade de 91.5% e uma especificidade de 34.2%.

O próximo passo consiste então em calcular o erro de teste para o ponto de corte escolhido.

O erro de teste é calculado da seguinte forma:

$$\text{Erro teste} = \frac{\sum \text{observações classificadas incorrectamente no conjunto de teste}}{\text{nº total de observações no conjunto de teste}}$$

Que no caso do modelo construído é:

$$= \frac{561666}{975147} = 0.576$$

Assim, prevê-se que quando se utilizar o modelo para classificar exemplos futuros, errar-se-á em 57.6% dos casos, o que é um erro grande mas que se deve principalmente ao facto do modelo ter apenas uma classificação razoável e ao facto de se preferir que o modelo acerte em grande parte das observações de clientes que anulam o contrato de seguro.

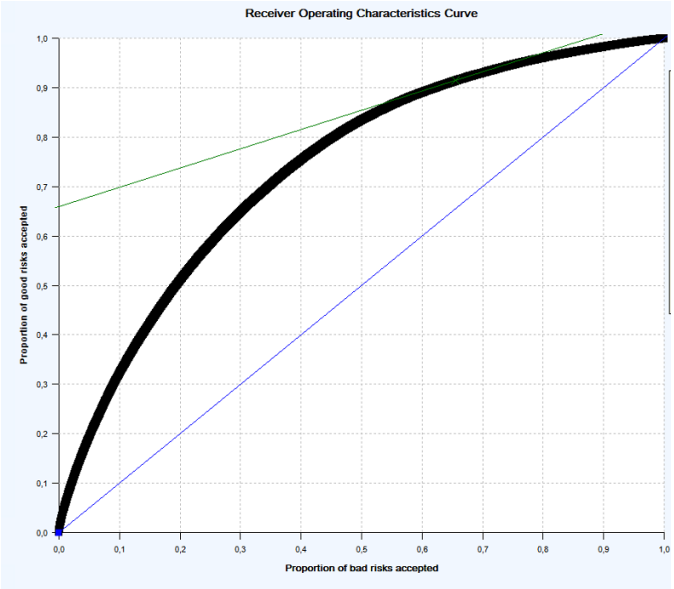


Figura 4. 3 - Curva ROC para o modelo com ponto de corte de 0.06.

Na figura 4.3 encontra-se representada a curva ROC a preto com o ponto de corte escolhido de 0.06 representado pela tangente que a curva verde faz com a curva.

A linha a azul representa o modelo que dá resultados positivos ou negativos aleatoriamente, ou seja, o modelo sem discriminação. Este serve de comparação ao nosso modelo construído. Como se pode ver, a curva ROC do nosso modelo não se aproxima da curva ROC do modelo aleatório, indicando que o nosso modelo possui alguma discriminação, mais precisamente AUC=0.7383, tendo portanto uma discriminação aceitável.

Note-se que no eixo das abcissas está representada a medida $(1 - \text{Especificidade})\%$ e no eixo das ordenadas a medida $\text{Sensibilidade}\%$.

4.5 Diagnósticos

Depois do modelo ter sido ajustado, é importante olhar para os gráficos de diagnósticos para a eventual deteção de “outliers”, pontos com “leverages” altas, pontos influentes, etc.

A seguir segue-se o gráfico produzido no programa Emblem da dispersão dos resíduos DR_i contra o valor previsto do preditor linear $\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$:

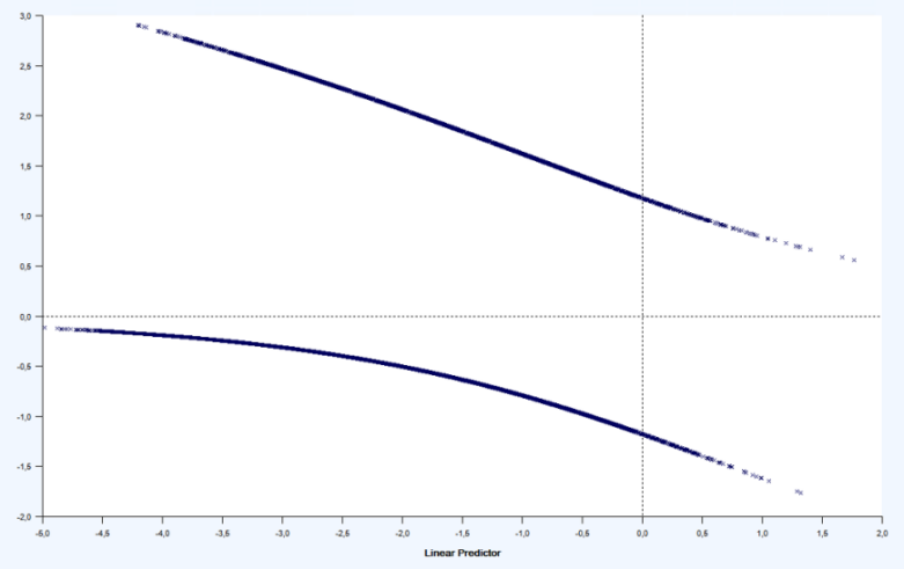


Figura 4. 4 - Resíduos estandardizados da desviância vs Preditor Linear

Como se pode ver pela figura 4.4, os resíduos estandardizados da desviância encontram-se em duas curvas de pontos. A curva acima corresponde à classe em que o cliente anula a apólice e a curva abaixo corresponde à classe em que o cliente não anula a apólice. Como na curva abaixo os valores absolutos dos resíduos são menores em relação aos da curva de cima então significa que a classe em que o cliente não anula a apólice está a ser melhor prevista. Contudo já tinha sido escolhido um ponto de corte na curva ROC de forma que a sensibilidade do modelo fosse elevada.

Espera-se que 95% dos resíduos estejam compreendidos entre -2 e 2, não existindo anomalias nesta análise se esta condição se verificar, como se pode constatar na figura 4.4.

Esta análise, em conjunto com a deteção de pontos influentes, é de extrema importância para a deteção de observações discordantes que podem enviesar o modelo.

Devido ao elevado número de dados, tornou-se computacionalmente muito exigente produzir quaisquer outros gráficos de diagnóstico no programa SAS. Além disso, no programa Emblem existem apenas disponíveis alguns gráficos de interesse como por exemplo, o da figura 4.4.

Assim sendo, apenas irei mencionar alguns gráficos relevantes [1]:

- Gráfico dos resíduos DR_i contra cada uma das variáveis explicativas do preditor linear. Se existir uma tendência, essa pode ser devida à escolha da função de ligação,

à escolha da escala usada numa ou mais variáveis explicativas ou à omissão de um termo quadrático na parte sistemática do modelo;

● Gráfico do preditor linear $\text{logit}(\hat{\pi}_i)$ contra as variáveis explicativas. Neste gráfico deve-se observar uma tendência linear;

● Gráfico das probabilidades observadas contra as probabilidades previstas. Aqui deve-se observar uma tendência linear, próxima da reta diagonal;

● Gráfico dos resíduos DR_i contra a ordem das observações. O gráfico não deve apresentar nenhuma tendência para que a hipótese de independência das observações seja satisfeita;

● O half-normal plot das “leverages”, das distâncias de Cook e dos resíduos DR_i. Os pontos influentes ou “outliers” aparecerão nos extremos dos gráficos e eventualmente sem seguir a tendência definida pelos restantes pontos. No caso das “leverages” e das distâncias de Cook esta tendência não é necessariamente linear;

● o qq-plot dos resíduos DR_i. Aqui é de esperar desvios em relação à normalidade, especialmente se o tamanho amostral for baixo.

Capítulo 5

Comparação dos Resultados com [25]

Como referido por [25], este estudo de modelação e previsão de anulações no seguro automóvel tem que ser revisto com alguma periodicidade, sendo que o objetivo principal desta dissertação foi essencialmente esse, o de atualizar o estudo com dados mais recentes e o de melhorar a construção do modelo final.

Neste sentido, irei realizar algumas comparações com o trabalho de [25] em termos da abordagem adotada e da modelação do conteúdo.

Enquanto [25], foca-se somente na análise descritiva das variáveis explicativas e na construção do modelo final, o trabalho apresentado nesta dissertação teve além disso, foco em alguns aspetos que podem comprometer o correto ajustamento do modelo final.

Em primeiro lugar, neste trabalho foram tidas em conta a avaliação da dependência estatística entre cada uma das variáveis e a resposta, a avaliação do nível de associação entre pares dessas variáveis explicativas, assim como o agrupamento de categorias com frequências próximas de zero em algumas das variáveis antes de ser feita a construção do modelo de regressão logística. Estas análises prévias à construção do modelo são importantes uma vez que evitam que variáveis redundantes sejam incluídas no modelo, assim como evitam que surjam coeficientes e/ou erros padrão demasiado elevados.

Para selecção do modelo final, a abordagem tomada por [25] foi a de criar um modelo final por tentativa e erro, incluindo ou não variáveis no modelo e comparando modelos com base nos valores de AUC e AIC. Já nesta dissertação, considerei no total 17 modelos, um com todas as variáveis (todas as variáveis que não foram consideradas redundantes), outro com todas as variáveis exceto “Semestre de Vencimento” e “Pack Automóvel” (por suspeita de não serem significativas para o modelo) e os restantes 15 modelos foram obtidos a partir do segundo mas retirando as 15 variáveis uma em cada modelo para assim poder comparar com o segundo modelo (uma vez que tinha sido decidido que este era melhor do que o primeiro).

Depois do modelo ter sido ajustado, algumas categorias não se revelaram significativas pelo que agrupei estas categorias não significativas a fim de evitar coeficientes e/ou erros padrão demasiado elevados, sendo que esta análise não foi considerada em [25].

Em ambos os trabalhos foi estudada a hipótese de inclusão de interações tendo-se chegado à conclusão que essas interações não melhoravam significativamente o modelo e portanto não deveria ser incluídas no modelo.

Nesta dissertação, foram realizados testes de hipóteses com o objetivo de avaliar a significância estatística dos coeficientes de regressão do modelo final sendo que estes não foram realizados em [25].

De seguida, foi escolhido um ponto de corte adequado na curva ROC e medida a capacidade de previsão do modelo, sendo que em [25] essa capacidade de previsão do modelo foi feita através de intervalos de confiança, não havendo referência do ponte de corte usado.

Finalmente, nesta dissertação foram mencionados alguns gráficos de diagnósticos importantes para deteção de “outliers”, pontos com “leverages” altas, pontos influentes, etc. No entanto esta análise não foi realizada dada a dificuldade computacional que exigia.

Concluindo, o objetivo de melhorar e atualizar o estudo feito por [25] foi atingido apesar dos modelos encontrados em ambos os relatórios terem uma classificação de razoável segundo o valor de AUC. É de notar também que os modelos obtidos nos dois trabalhos são diferentes devido às diferentes abordagens adotadas em ambos os trabalhos.

Capítulo 6

Conclusões e trabalho futuro

O propósito deste trabalho centrou-se no estudo do modelo de regressão logística e na sua aplicação a dados sobre anulações de contratos no seguro automóvel. Mais concretamente, a Ageas colocou-me o problema da identificação de variáveis que levem a que o cliente anule o seu contrato de seguro automóvel. As variáveis que mostraram ter um efeito significativo sobre a resposta foram o número de sinistros, a existência ou não de outras apólices na seguradora, a antiguidade do contrato, o escalão de bónus, a evolução no escalão de bónus, o sexo do tomador do seguro, o intervalo estimado de variação do prémio, a região onde reside o tomador do seguro, a forma de cobrança, a forma de pagamento, a categoria automóvel, a idade do condutor, o tempo de carta, a rede e o produto.

Houve no entanto uma limitação ao nível de capacidade computacional do programa SAS que não permitiu a produção de gráficos de diagnósticos importantes para a deteção de observações discordantes que podem enviesar o modelo como por exemplo, o gráfico dos resíduos DR_i contra cada uma das variáveis explicativas do preditor linear, o gráfico do preditor linear $\mathrm{logit}(\hat{\pi}_i)$ contra as variáveis explicativas, o gráfico das probabilidades observadas contra as probabilidades previstas, o gráfico dos resíduos DR_i contra a ordem das observações, o half-normal plot das “leverages”, das distâncias de Cook e dos resíduos DR_i e o qq-plot dos resíduos DR_i.

Como trabalho futuro, sugiro a exploração de novas variáveis que consigam melhorar a capacidade explicativa do modelo uma vez que o modelo encontrado tem apenas uma discriminação aceitável. É de notar no entanto que a oferta feita aos clientes por outras seguradoras influencia igualmente a decisão de escolha por parte destes, o que torna a tarefa de previsão de anulações ainda mais difícil.

Em suma, apesar das dificuldades computacionais encontradas, considero que o estudo realizado foi positivo para a AGEAS e permitirá à seguradora prever o comportamento dos seus clientes face a certas situações e assim conseguir diminuir a taxa de anulações no seguro automóvel.

Referências

[1] A. Rita Gaio, Apontamentos escritos da disciplina EACE sobre Regressão Logística: Mestrado em Engenharia Matemática, Departamento de Matemática FCUP, 2013

[2] A. Rita Gaio, Apontamentos escritos da disciplina Estatística Aplicada sobre Testes de Hipóteses: Dados Categóricos: Licenciatura em Matemática, Departamento de Matemática FCUP, 2011

[3] P. David Allison, Logistic Regression Using the SAS System: Theory and Application, 2nd edition, University of Pennsylvania, 2012

[4] R. Christensen, Log-Linear Models and Logistic Regression, 2nd edition, Springer Texts in Statistics, 1997

[5] G. Tutz, Regression for Categorical Data, Cambridge Series in Statistical and Probabilistic Mathematics, 2012

[6] D. W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd edition, Wiley Series in Probability and Statistics, 2000

[7] G. Rodriguez, Universidade de Princeton, apontamentos online, http://data.princeton.edu/wws509/

[8] J. W. Hardin, J. m. Hilbe, Generalized Linear Models and Extensions, 2nd edition, A Stata Press Publication, 2001

[9] A. Agresti, An Introduction to Categorical Data Analysis, 2nd edition, John Willey & Sons, INC, 1996

[10] D. G. Kleinbaum, M. Klein, Logistic Regression: A Self-Learning Text, 3rd edition, Emory University, 2010

[11] Tomas W. Yee, Vector Generalized Linear and Additive Models: With an Implementation in R, Springer Series in Statistics, 2015

[12] G. M. Fitzmaurice, N. M. Laird, J. H. Ware, Applied Longitudinal Analysis, 2nd edition, Wiley Series in Probability and Statistics, 2011

[13] S. Menard, Logistic Regression: From introductory to advanced concepts and applications, Sam Houston State University, 2009

[14] A. Agresti, Categorical Data Analysis, 2nd edition, John Willey & Sons, INC, 2002

[15] T. Fawcett, An Introduction to ROC analysis, Pattern Recognition Letters 27: 861-874, 2006

[16] S. Menard, Applied Logistic Regression Analysis, 2nd edition, Series: Quantitative Applications in the Social Sciences, 2002

[17] P. McCullagh, J. A. Nelder, Generalized Linear Models, 2nd edition, Monographs on Statistics and Applied Probability 37, 1989

[18] J. M. Hilbe, A. P. Robinson, Methods of Statistical Model Estimation, A Chapman & Hall Book, 2013

[19] J. M. Hilbe, Pratical Guide to Logistic Regression, A Chapman & Hall Book, 2015

[20] M. Refaat, Data Preparation for Data Mining Using SAS, Morgan Kaufmann Publishers, 2007

[21] D. Collet, Modeling Binary Data, 2nd edition, Texts in Statistical Science, Chapman & Hall/CRC, 2003

[22] E. W. Frees, R. A. Derrig, G. Meyers, Predictive Modeling Applications in Actuarial Science: Volume I: Predictive Modeling Techniques, International Series on Actuarial Sciences, 2014

[23] R. M. Warner, Applied Statistics: From Bivariate Through Multivariate Techniques, 2nd edition, SAGE Publications, Inc., 2013

[24] http://groups.chass.utoronto.ca/pol242/Labs/LM-3A/LM-3A_content.htm

[25] J. Rita Barradas Garraio, Modelação da Taxa de Anulação no Seguro Automóvel, Departamento de Estatística e Investigação Operacional FCUL, 2015

Anexos

Anexo A

Gráficos de percentagens de anulações de apólices por valor de cada variável

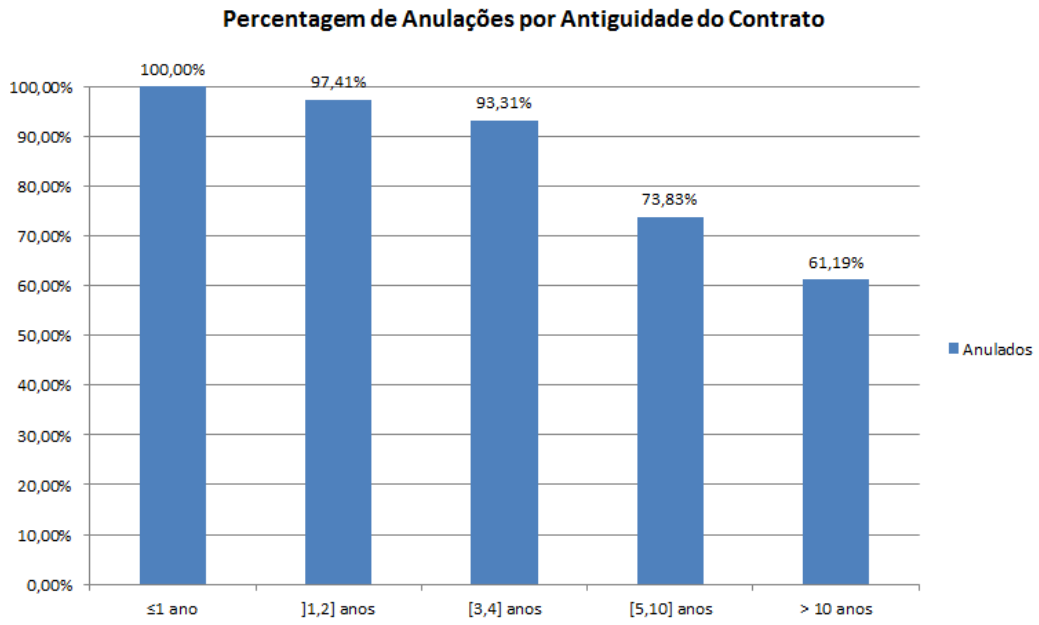


Figura A. 1 - Anulações vistas pelos valores da variável Antiguidade do Contrato

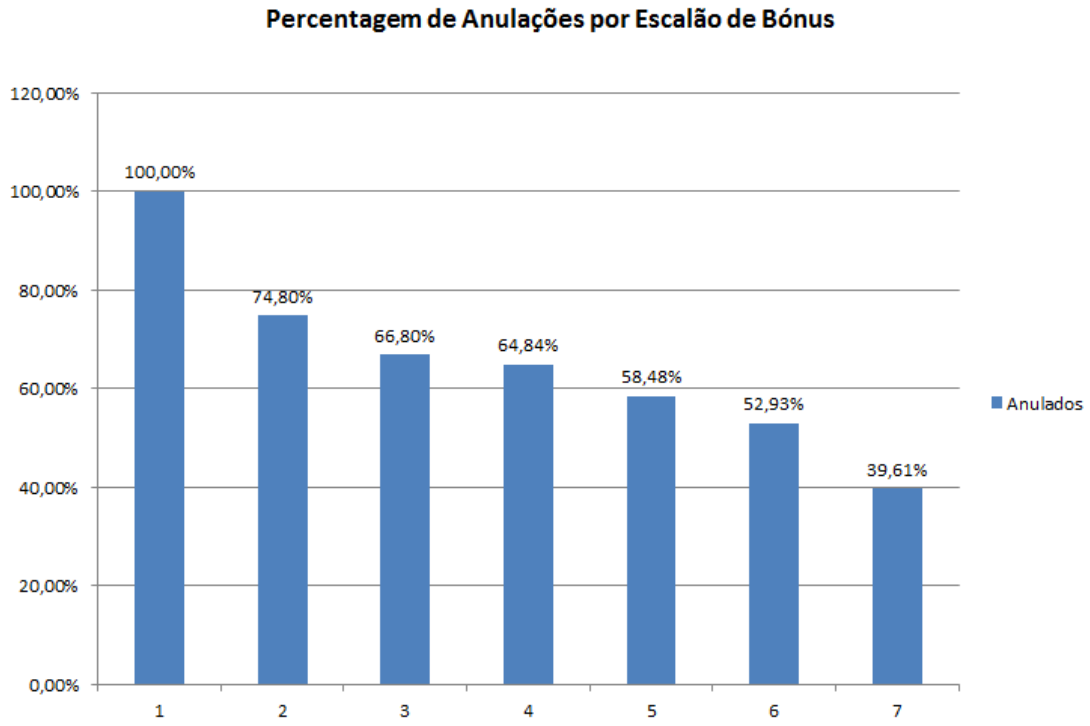


Figura A. 2 - Anulações vistas pelos valores da variável Escalão de Bónus

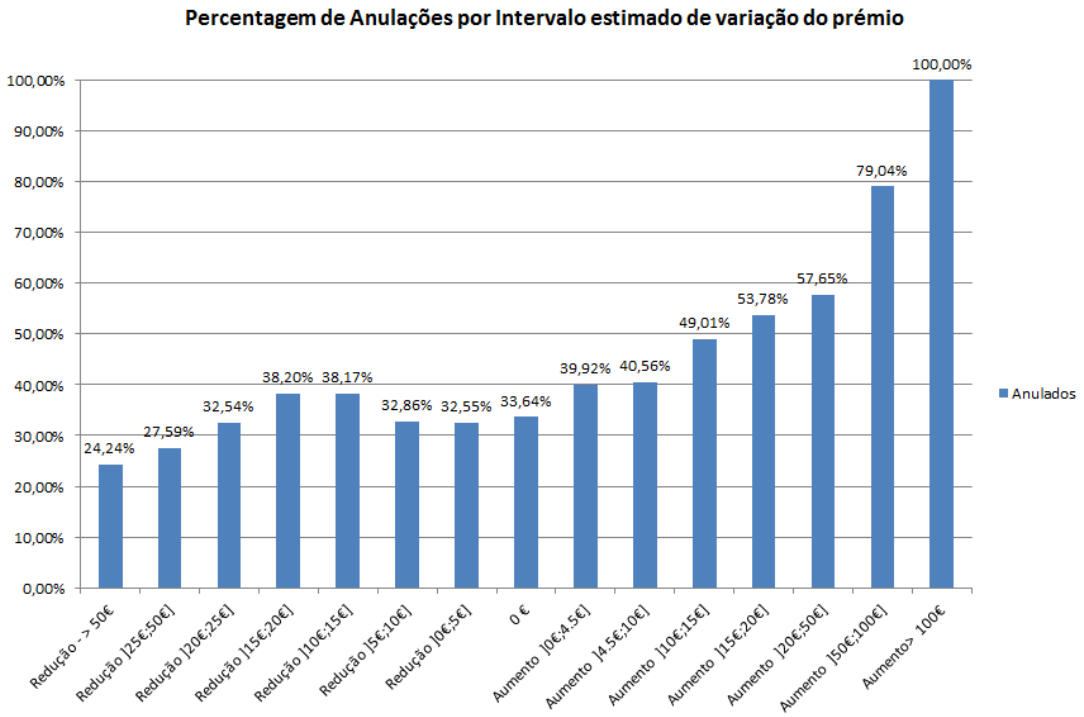


Figura A. 3 - Anulações vistas pelos valores da variável Intervalo estimado de variação do prémio

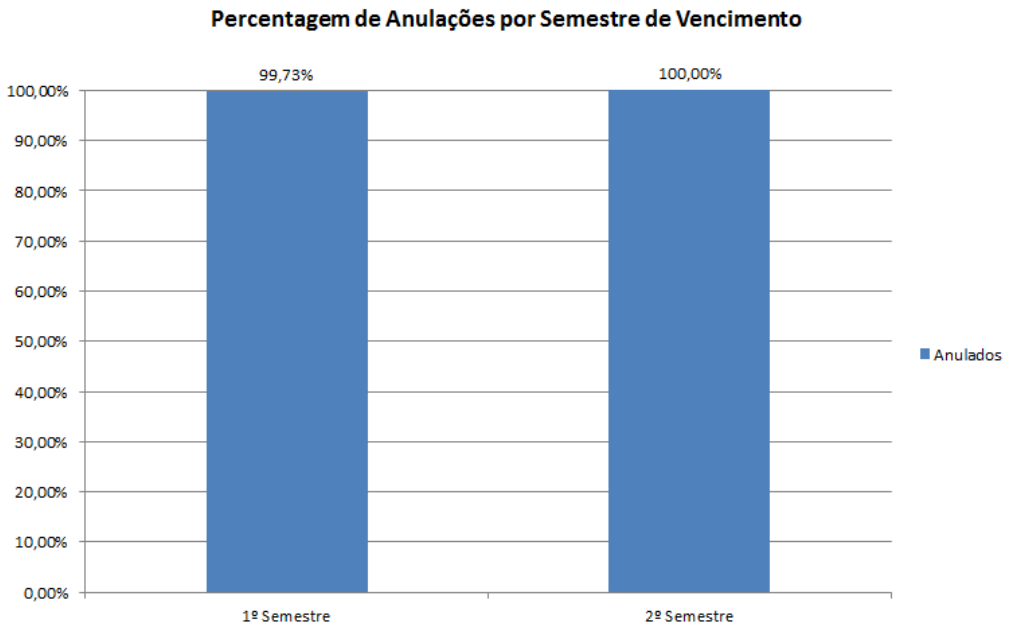


Figura A. 4 - Anulações vistas pelos valores da variável Semestre de Vencimento

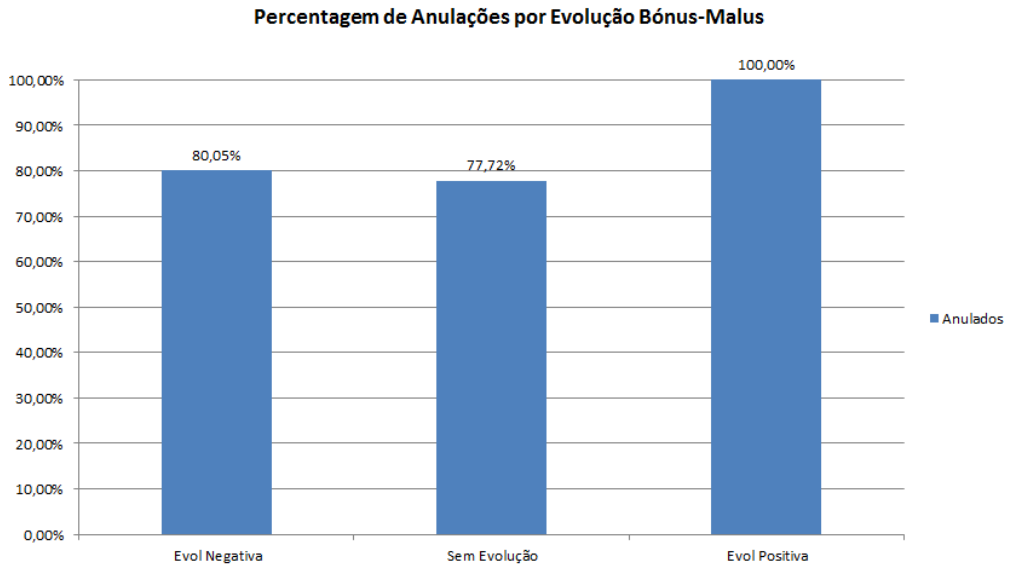


Figura A. 5 - Anulações vistas pelos valores da variável Evolução Bónus-Malus

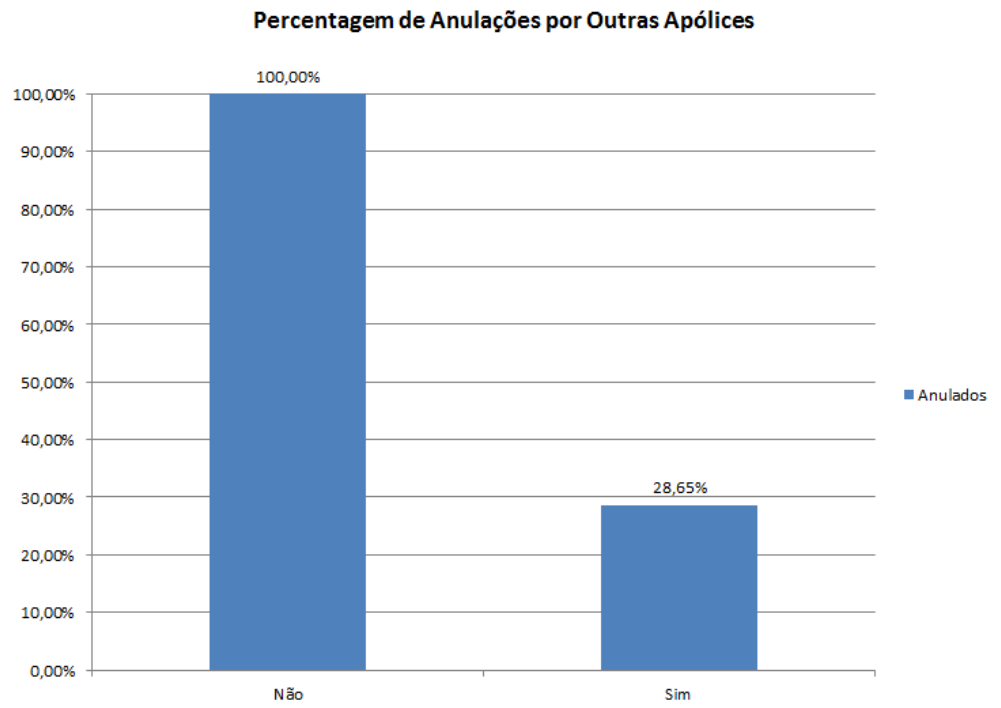


Figura A. 6 - Anulações vistas pelos valores da variável Outras Apólices

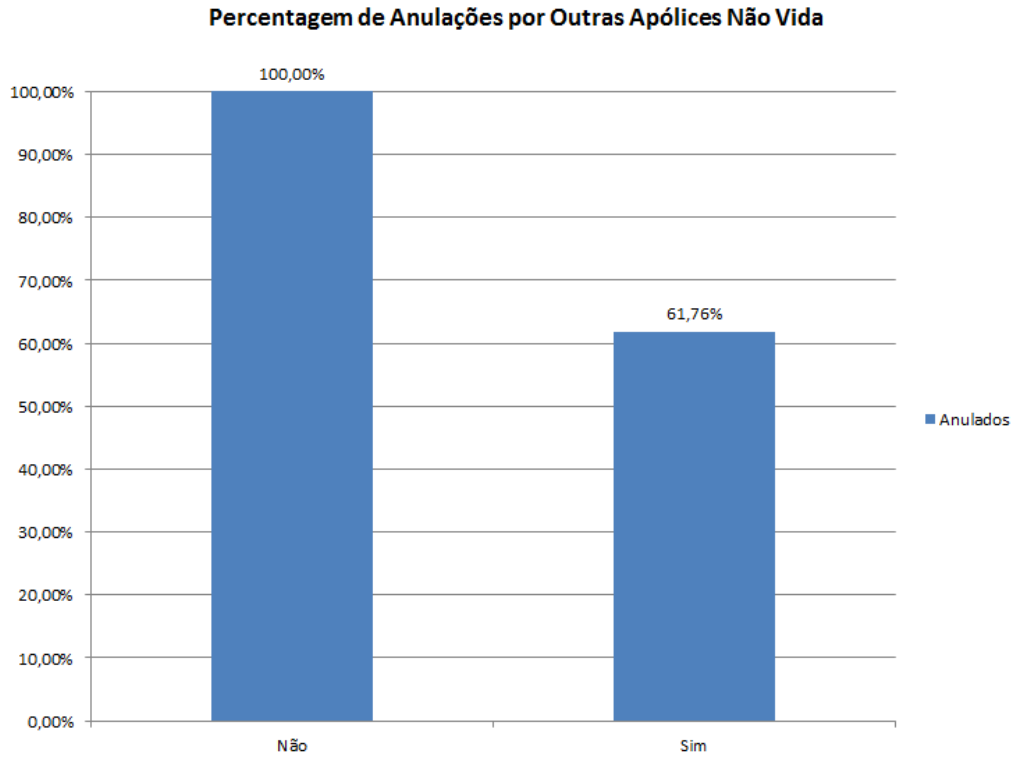


Figura A. 7 - Anulações vistas pelos valores da variável Outras Apólices Não Vida

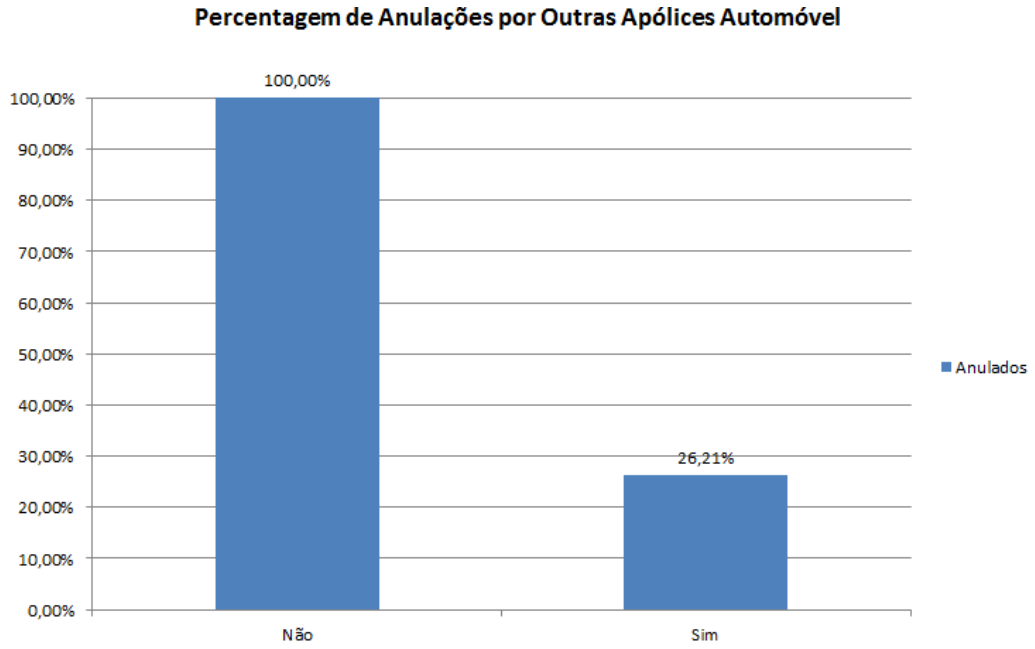


Figura A. 8 - Anulações vistas pelos valores da variável Outras Apólices Automóvel

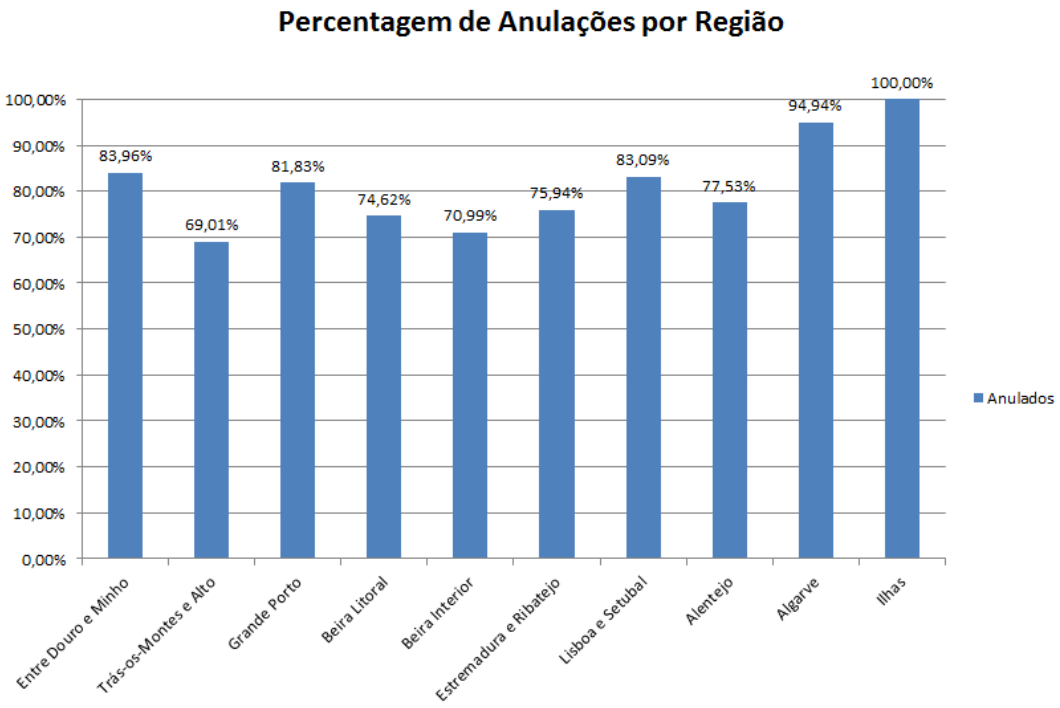


Figura A. 9 - Anulações vistas pelos valores da variável Região

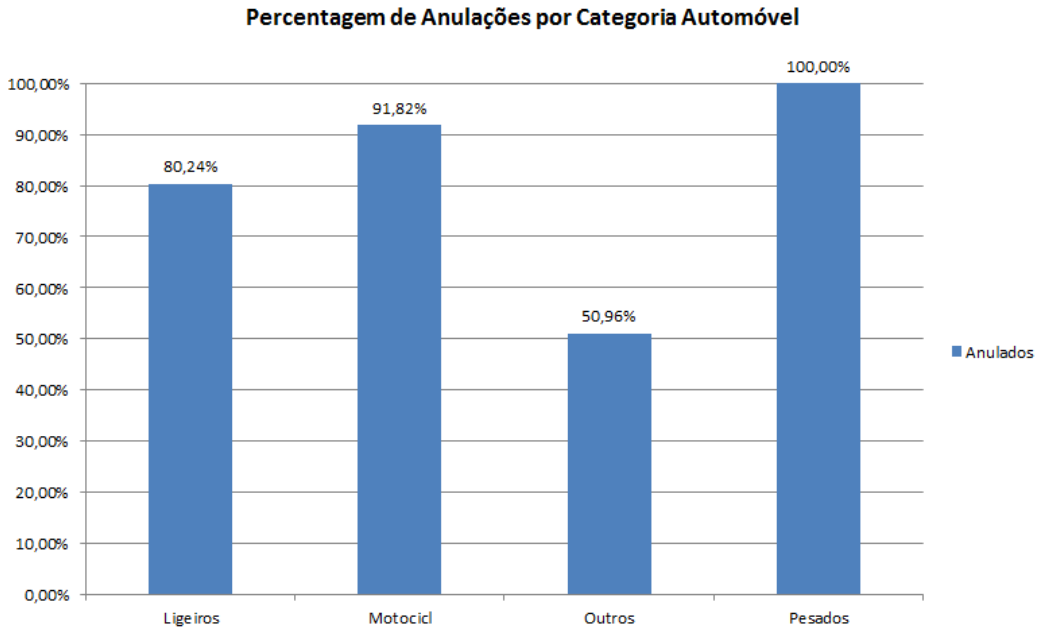


Figura A. 10 - Anulações vistas pelos valores da variável Categoria Automóvel

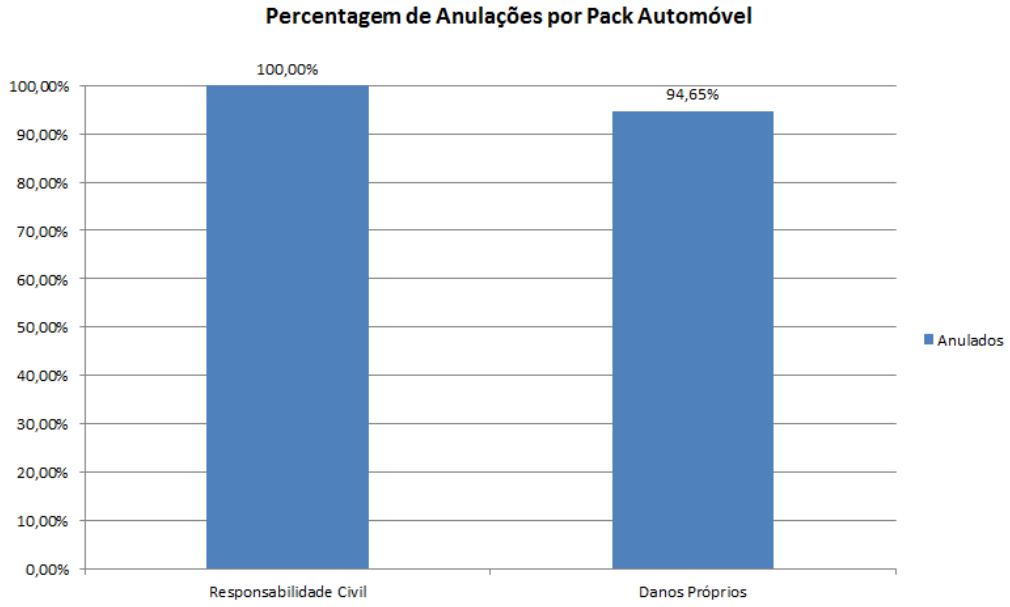


Figura A. 11 - Anulações vistas pelos valores da variável Pack Automóvel

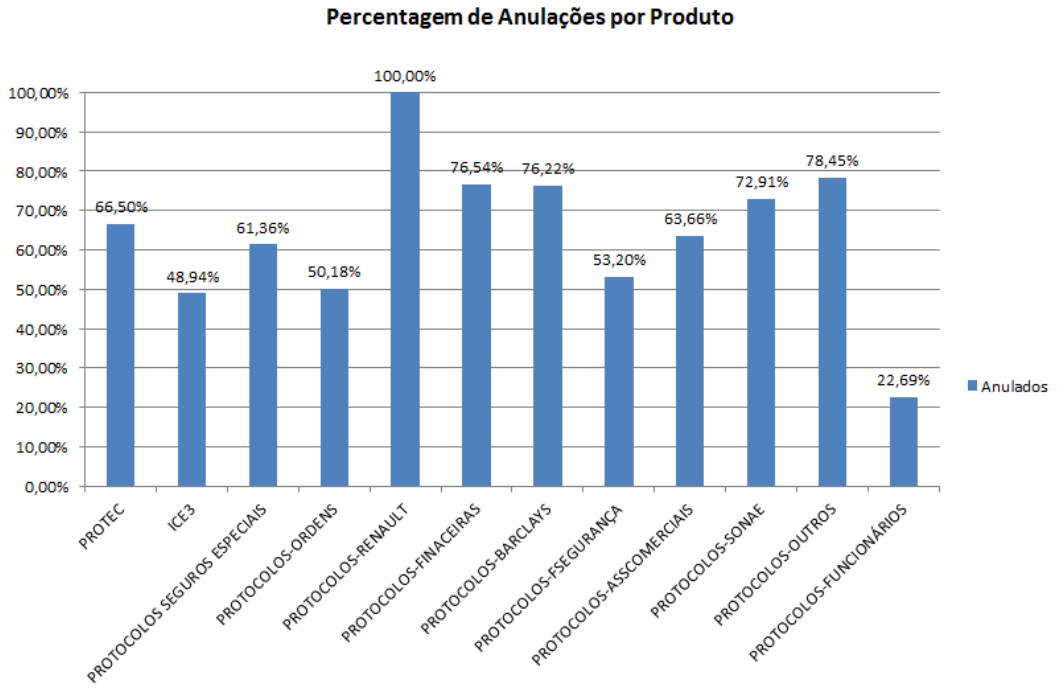


Figura A. 12 - Anulações vistas pelos valores da variável Produto

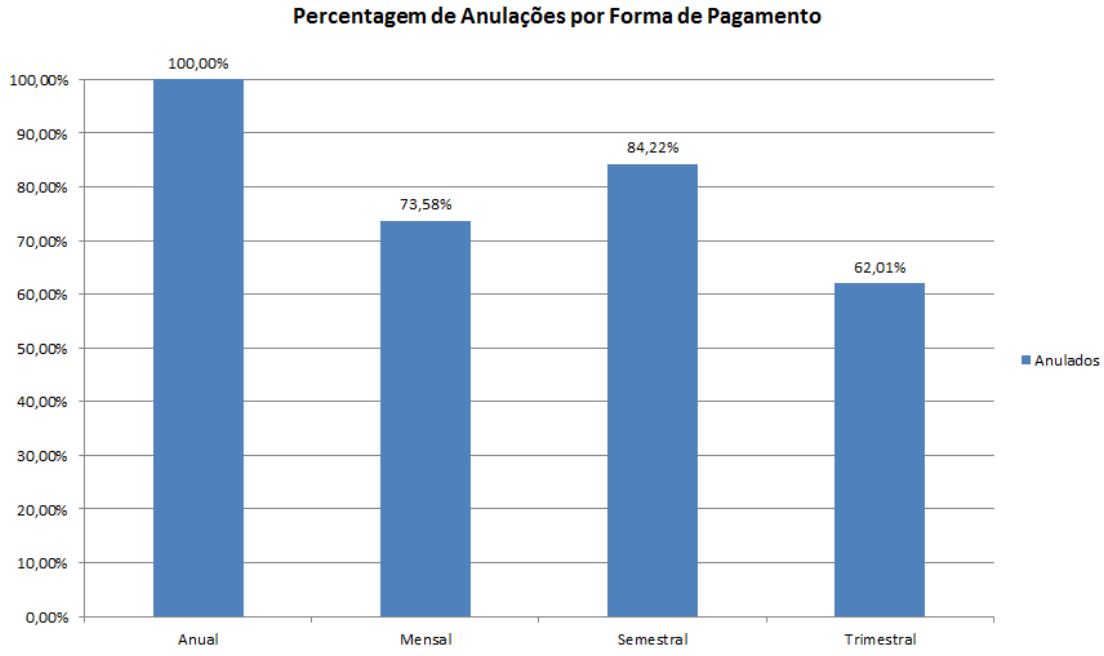


Figura A. 13 - Anulações vistas pelos valores da variável Forma de Pagamento

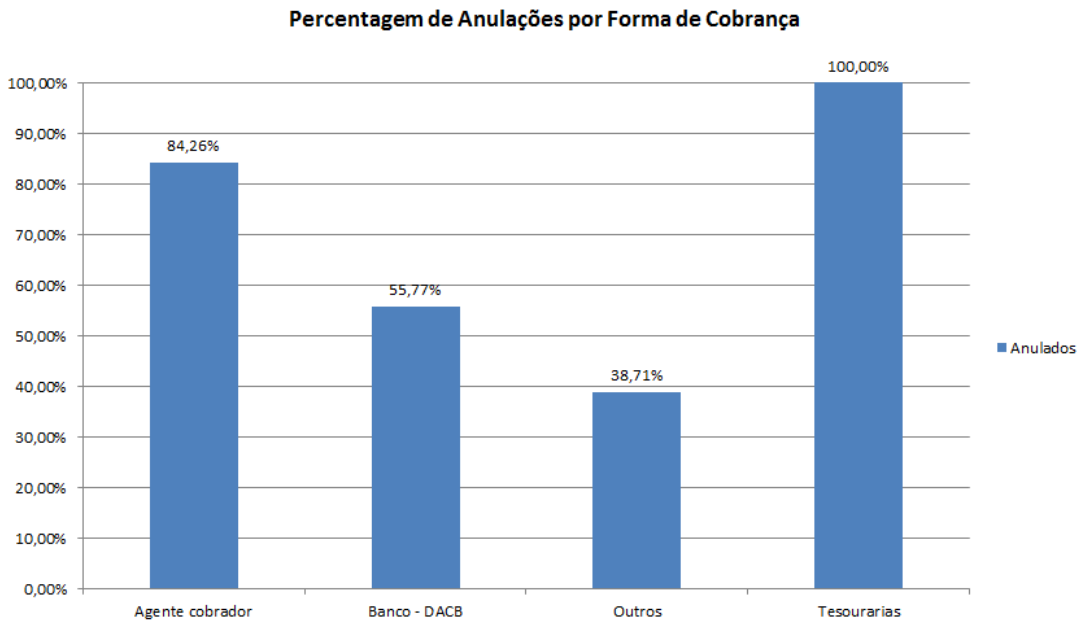


Figura A. 14 - Anulações vistas pelos valores da variável Forma de Cobrança

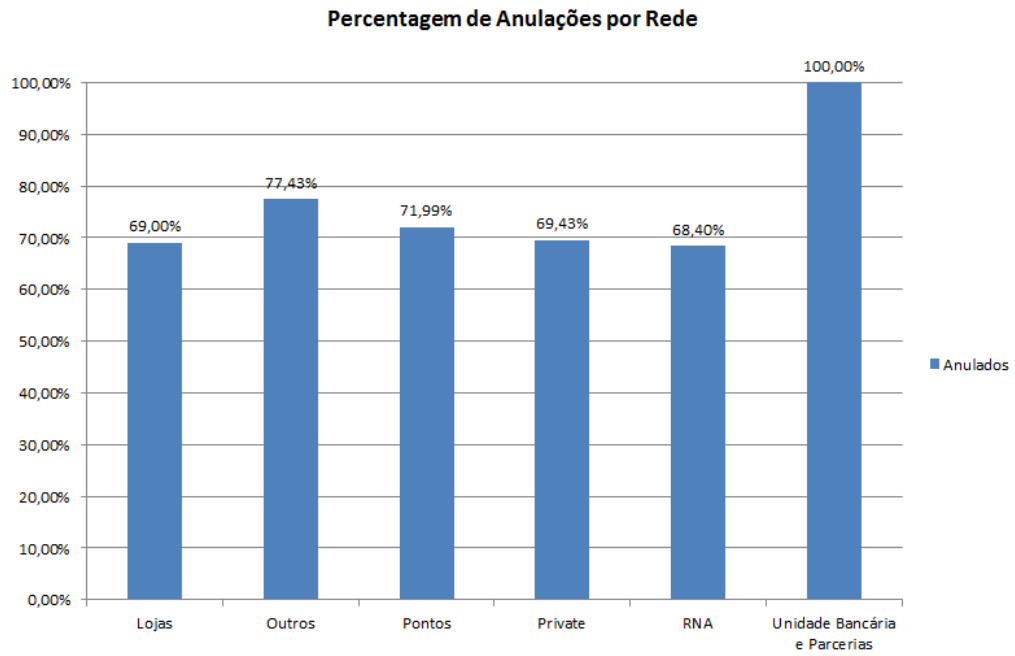


Figura A. 15 - Anulações vistas pelos valores da variável Rede

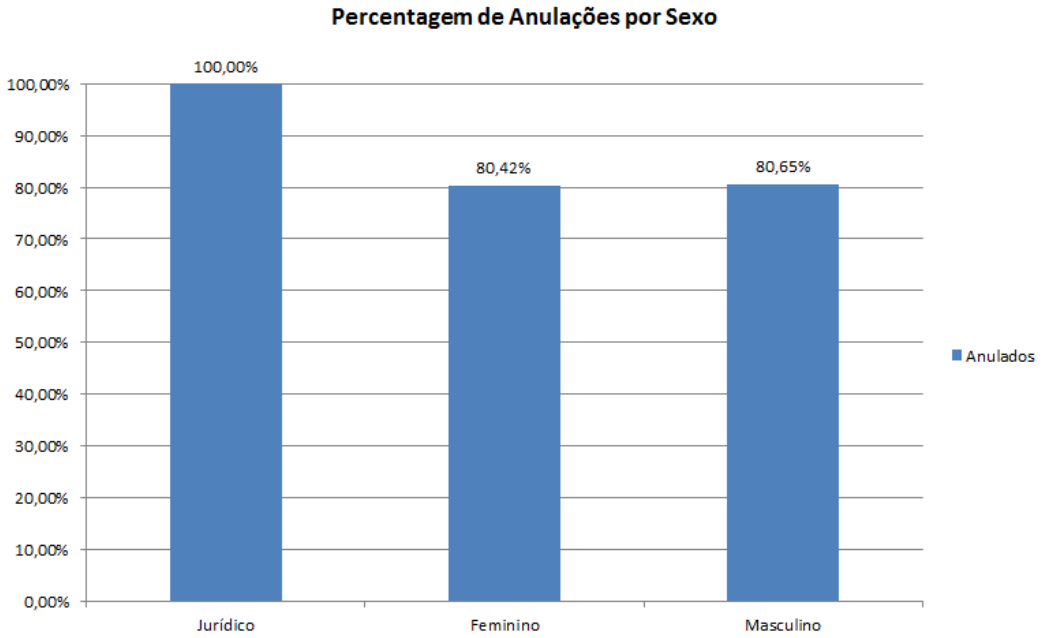


Figura A. 16 - Anulações vistas pelos valores da variável Sexo

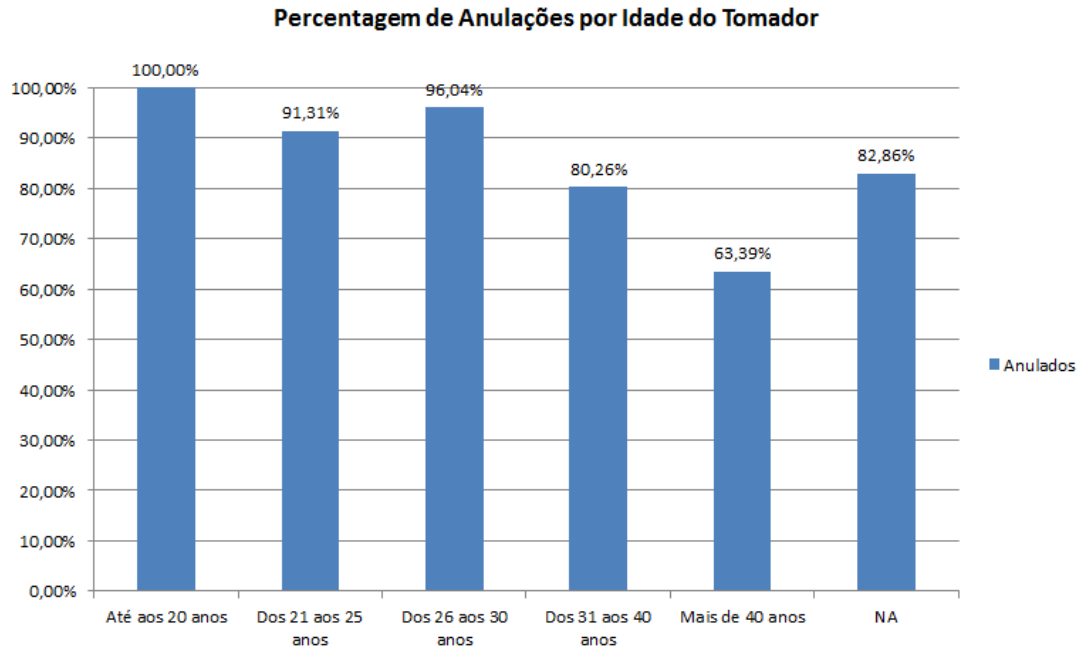


Figura A. 17 - Anulações vistas pelos valores da variável Idade do Tomador

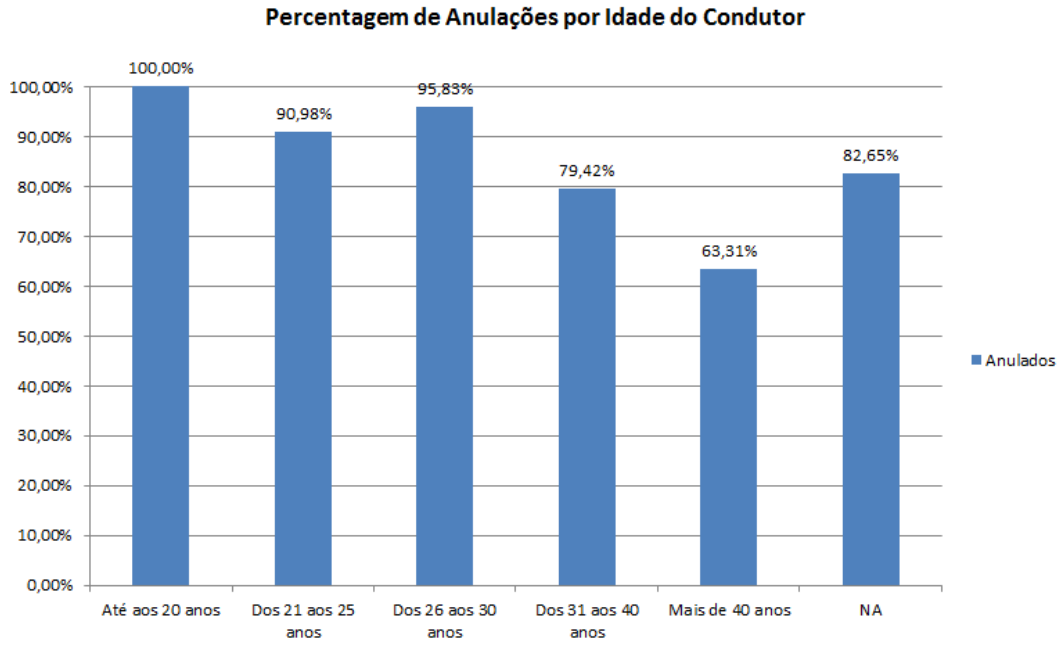


Figura A. 18 - Anulações vistas pelos valores da variável Idade do Condutor

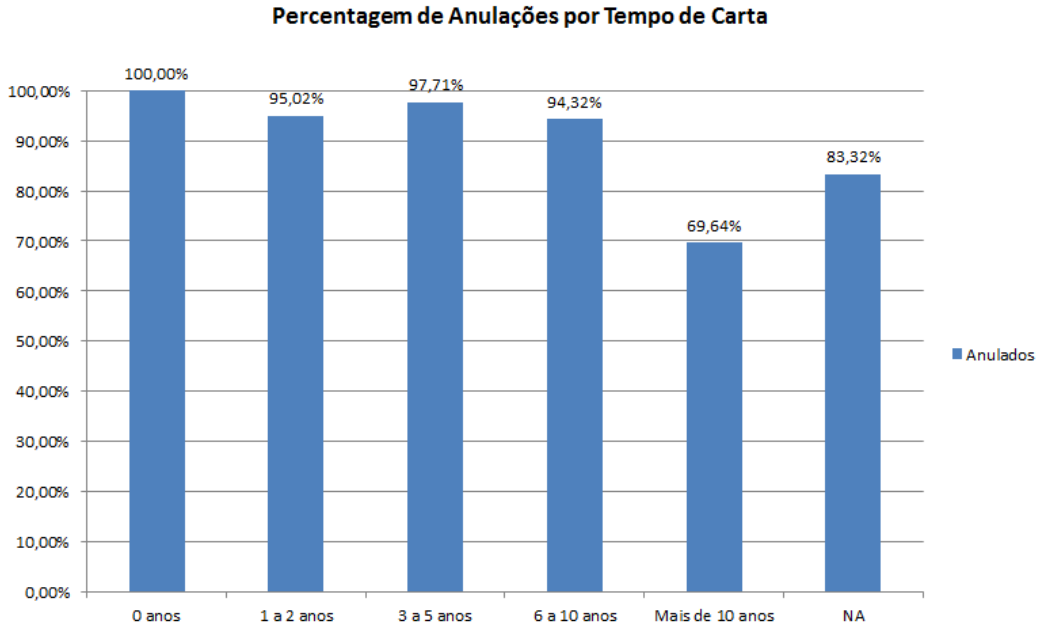


Figura A. 19 - Anulações vistas pelos valores da variável Tempo de Carta

Anexo B

Código SAS

B.1 Construção do modelo de Regressão Logística

```
libname tabsas 'C:\Beatriz\Relatório de Estágio\Beatriz\sas\output sas
final';
libname export 'X:\';

data final;
set tabsas.final_corrigida2;
run;

/*Teste de independência entre Codanula e Sinistralidade*/
proc freq data=final;
 tables codanula*tevesin / chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Antiguidade do Contrato*/
proc freq data=final;
 tables codanula*tempocont / chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Escalão de Bónus*/
proc freq data=final;
 tables codanula*escaloes / chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Semestre de Vencimento*/
proc freq data=final;
 tables codanula*semestre/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Pack Automóvel*/
proc freq data=final;
 tables codanula*packAuto/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;
```

```
/*Teste de independência entre Codanula e Intervalo Estimado de
Variação do Prémio*/
proc freq data=final;
 tables codanula*int_estimado/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Evolução Bónus-Malus*/
proc freq data=final;
 tables codanula*evoBonusMalus/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Outras Apólices*/
proc freq data=final;
 tables codanula*outrasApos/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Outras Apólices Não Vida*/
proc freq data=final;
 tables codanula*outrasAposNV/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Outras Apólices Automóvel*/
proc freq data=final;
 tables codanula*outrasAposAuto/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Região*/
proc freq data=final;
 tables codanula*regiao/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Categoria Automóvel*/
proc freq data=final;
 tables codanula*categor2/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Produto*/
proc freq data=final;
 tables codanula*produto_aux/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Forma de Pagamento*/
proc freq data=final;
 tables codanula*fPagamento/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;
```

```
/*Teste de independência entre Codanula e Forma de Cobrança*/
proc freq data=final;
 tables codanula*fcobranca/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Rede*/
proc freq data=final;
 tables codanula*d_rede/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Sexo*/
proc freq data=final;
 tables codanula*sexo/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Idade do Tomador*/
proc freq data=final;
 tables codanula*idadetomador/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Idade do Condutor*/
proc freq data=final;
 tables codanula*idadecondutor/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;

/*Teste de independência entre Codanula e Tempo de Carta*/
proc freq data=final;
 tables codanula*TempoCarta/ chisq measures
 plots=(freqplot(twoway=groupvertical scale=percent));
run;
/*    VARIÁVEIS    EXPLICATIVAS:semestre,    int_estimado,    escaloes,
evoBonusMalus,   regiao,   outrasApos,   outrasAposNV,   outrasAposAuto,
categor2, packAuto, produto, fPagamento,fcobranca, tempocont, tevesin,
d_rede, sexo, idadecondutor, idadetomador, TempoCarta*/

/*CONJUNTO DE TREINO*/
data tabsas.conjunto_treino;
set     final(keep=New_d_rede     New_produto_aux_2     New_idadecondutor
New_int_estimado_2
New_produto_aux    New_int_estimado    regiao_aux    produto_aux    ano2
tempocarta2 idadetomador2
idadecondutor2   int_estimado_2   ano   codanula   cod_vigor   semestre
int_estimado escaloes
evoBonusMalus  regiao  outrasApos  outrasAposNV  outrasAposAuto  categor
categor2  packAuto
produto   fPagamento   fcobranca   tempocont   tevesin   d_rede   sexo
idadecondutor idadetomador
seg TempoCarta);
call streaminit(777);
  u = rand("Uniform");
  if u<=0.7;
run;
```

```
/*CONJUNTO DE TESTE*/
data tabsas.conjunto_teste;
set final (keep=New_d_rede New_produto_aux_2 New_idadecondutor
New_int_estimado_2
New_produto_aux New_int_estimado regiao_aux produto_aux ano2
tempocarta2 idadetomador2
idadecondutor2 int_estimado_2 ano codanula cod_vigor semestre
int_estimado escaloes
evoBonusMalus regiao outrasApos outrasAposNV outrasAposAuto categor
categor2 packAuto
produto fPagamento fcobranca tempocont tevesin d_rede sexo
idadecondutor idadetomador
seg TempoCarta);
call streaminit(777);
  u = rand("Uniform");
  if u>0.7;
run;

/*Para modelar os 1's em vez dos 0's, usa-se a opção descending.
Isto porque por defeito o proc logistic modela os 0's em vez dos
1's.*/
/*A opção aggregate faz com que os dados sejam agrupados*/
title "Modelo de Regressão Logística";
ods graphics on;
proc logistic data=tabsas.conjunto_treino descending
plots(maxpoints=NONE)
outmodel=tabsas.modeltreino;
class New_idadecondutor(ref="4. Mais de 40")
New_int_estimado_2(ref="05-0 €")
New_produto_aux_2(ref="01-PROTEC") sexo(ref="2")
outrasApos(ref="SIM")
fPagamento(ref="Anual") escaloes(ref="7")fcobranca(ref="Agente
cobrador")
categor2(ref="Ligeiros") evoBonusMalus(ref="1. Sem Evolução")
tevesin(ref="Não teve Sin ultm. 5 anos.") tempocarta(ref="Mais de 10
anos")
regiao_aux(ref="04. Beira Litoral")
tempocont(ref="01. Até 1 ano") New_d_rede(ref="RNA")/param=ref;
model codanula= New_idadecondutor New_int_estimado_2 New_produto_aux_2
sexo
outrasApos fPagamento escaloes fcobranca categor2 evoBonusMalus
tevesin
tempocarta regiao_aux tempocont New_d_rede / aggregate scale=none covb
lackfit rsquare
 outroc=troc roceps=0 ctable pprob=0.01 0.02 0.03 0.04 0.05 0.06 0.07
0.08 0.09 0.1 0.2
0.3 0.4 0.5 0.6 0.7 0.8 0.9;
output out=results p=yhat ;
         ods output  classlevelinfo=dum;
         ods output  association=assoc;
         ods output parameterestimates = estimates_par CovB=lgscovb;
         ods output oddsratios = estimates_int;
      run;
```

```
/*Comandos para escolher o ponto de corte para o modelo de treino*/
data results;
set results;
if yhat ge .06 then codanula_previsto=1;
else codanula_previsto=0;
run;

/*Tabela com os valores observados e previstos parao ponto de
corte escolhido anteriormente*/
proc freq data=results;
table codanula*codanula_previsto/ nopercent norow nocol;
run;


/*Neste comando é calculado o erro do modelo: I_codanula são as
previsões
do modelo para o conjunto de teste e F_codanula são os valores
observados
do conjunto de teste*/
/*No entanto este erro é calculado para um ponto de corte=0.5...*/
proc logistic inmodel=tabsas.modeltreino;
score      data=tabsas.conjunto_teste      out=test_scored      fitstat
outroc=roc_teste ;
ods output association=assoc;
run;

/*Neste comando é apresentada a tabela de contigência e a partir dela
podemos calcular o erro no conjunto de teste para um ponto de corte de
0.5...*/
proc freq data=test_scored;
      table F_codanula*I_codanula / nocol nocum nopercent;
   run;


/*Para escolher outro ponto de corte por exemplo:0.06 faz-se o
seguinte:*/
data a;
set test_scored;
if P_1<'0.06' then I_codanula='0';
if P_1>='0.06' then I_codanula='1';
keep P_0 P_1 I_codanula F_codanula;
run;

/*E a partir desta tabela pode-se calcular o erro de teste para o
ponto
de corte de 0.06*/
proc freq data=a;
      table F_codanula*I_codanula / nocol nocum nopercent;
         run;
```

B.2 Passagem do modelo de Regressão Logística para Emblem

```
libname guarda 'C:\Beatriz Faustino';
libname Emblem 'C:\Beatriz Faustino\Emblem treino todas as variáveis';

options fmtsearch = (guarda);*Asignamos la librería de los formatos de
las variables;


%let  Emblem=C:\Beatriz  Faustino\Emblem  treino  todas  as  variáveis;
*Ubicación del fichero output de Emblem;


*********************************************************************
************************************************************
CREAR ARCHIVOS EMBLEM
*********************************************************************
*****************************************************;

%let factores=
outrasAposNV
outrasAposAuto
packAuto
idadetomador
OutrasApos
sexo
int_estimado
escaloes
fPagamento
fcobranca
produto_aux
categor2
teveSin
TempoCont
idadecondutor
d_rede
regiao_aux
TempoCarta
evoBonusMalus
semestre
/*seg*/
/*ano*/;

data anula;
set guarda.conjunto_treino;
apol=codanula+cod_vigor;
run;
```

```
%PrepFile(  Dataset =anula,
                  Factors = &factores,
                  Store = emblem.anula_s,
                  Variates = ,
                  Fmtvalue = all);


%macro crear_modelos_emblem (bdstore,bbdd,modelo,apol,codanula);

*crea el "store" que es una lista de factores de riesgo y sus niveles;
*a partir de una bbdd y una lista de factores;
*cada factor tiene menor o igual a 255 niveles;
*fmtvalue = yes coge como niveles los valores de un formato en lugar
de los valores subyacentes;


*esto crea un fichero emblem;
*te pide un store y una bbdd;
*esta bbdd no tiene porque ser la misma de antes, pero en general será
un subconjunto de la bbdd
 utilizada para crear el store;

%CrteFile(  Store=emblem.&bdstore,
                  Dataset=&bbdd, /*nombre base de datos (más abajo)*/
                  DirName=&Emblem, /*path directiorio (más arriba)*/
                  Title=&modelo, /*Nombre ficheros .bid y .fac*/
                  Weight=&apol, /*nombre variable expo (más abajo)*/
                  Response=&codanula, /*nombre variable respuesta (más
abajo)*/
                  Binomial=);


%mend;


%crear_modelos_emblem (anula_s,anula,tx_anula,apol,codanula);
```

Anexo C

Estimativa para os coeficientes das variáveis explicativas

Variável	Valores	$\hat{\beta}$	Desvio Padrão
Sinistralidade	Teve sinistros	0.2618	0.00698
	Não teve sinistros	0	0
Outras Apólices	Sim	0	0
	Não	1.6736	0.00489
Antiguidade do Contrato	≤ 1 ano	0	0
	]1,2] anos	0.0363	0.00671
	[3,4] anos	0.0563	0.00603
	[5,10] anos	-0.1035	0.00611
	> 10 anos	-0.1852	0.00789
Escalão de Bónus	1	0.7540	0.0285
	2	0.6466	0.0121
	3	0.5111	0.0132
	4	0.3460	0.0106
	5	0.2269	0.00661
	6	0.1456	0.00582
	7	0	0
Evolução Bónus-Malus	Evolução Negativa	-0.1183	0.00742
	Sem Evolução	0	0
	Evolução Positiva	0.0203	0.00732
Sexo	Jurídico (se o cliente é empresa)	0.1176	0.0150
	Feminino	-0.2711	0.00500
	Masculino	0	0

Quadro C. 1 - Estimativa dos coeficientes das variáveis explicativas

Variável	Valores	$\widehat{\beta}$	Desvio Padrão
Intervalo Estimado de Variação do Prémio	Redução >50€	-0.4287	0.0318
	Redução ∈]20€;50€]	-0.2488	0.0217
	Redução ∈]15€;20€]	0.0916	0.0115
	Redução ∈]5€;15€]	-0.0402	0.0113
	Redução ∈]0€;5€]	-0.0561	0.0118
	0€	0	0
	Aumento ∈]0€;4.5€]	0.0558	0.00693
	Aumento ∈]4.5€;10€]	0.1041	0.00681
	Aumento ∈]10€;15€]	0.3285	0.00692
	Aumento ∈]15€;20€]	0.4727	0.00965
	Aumento ∈]20€;50€]	0.5578	0.00922
	Aumento ∈]50€;100€]	0.8688	0.0165
	Aumento> 100€	1.2155	0.0207
Região	Entre Douro e Minho	0.0492	0.00764
	Trás-os-Montes e Alto	0.1011	0.0114
	Grande Porto	0.0585	0.00635
	Beira Litoral	0	0
	Beira Interior	0.0985	0.0113
	Estremadura e Ribatejo	0.0825	0.00723
	Lisboa	0.0558	0.00722
	Setúbal	0.1251	0.0108
	Alentejo	0.1669	0.0111
	Algarve	0.2155	0.00948
	Ilhas	0.1080	0.0115
Forma de Cobrança	Agente Cobrador	0	0
	Banco – DACB	-0.3052	0.00733
	Outros	-0.6540	0.0573
	Tesourarias	0.0984	0.00770

Quadro C.1 – Continuação da página anterior

Variável	Valores	$\widehat{\beta}$	Desvio Padrão
Forma de Pagamento	Anual	0	0
	Semestral	-0.4415	0.00588
	Trimestral	-0.7905	0.0119
	Mensal	-0.5096	0.0140
Categoria Automóvel	Ligeiros	0	0
	Motociclos	0.1848	0.00734
	Outros	-0.2364	0.0141
	Pesados	0.2065	0.0166
Idade do Condutor	Até aos 25 anos	0.0634	0.0164
	Dos 26 aos 30 anos	0.1155	0.0105
	Dos 31 aos 40 anos	0.0692	0.00547
	Mais de 40 anos	0	0
	NA (se o cliente é empresa)	0.8927	0.0298
Tempo de Carta	0 anos	-0.4400	0.0640
	1 a 2 anos	-0.2406	0.0223
	3 a 5 anos	-0.1395	0.0151
	6 a 10 anos	-0.0785	0.00943
	Mais de 10 anos	0	0
	NA (se o cliente é empresa)	-0.4391	0.0263
Rede	Lojas	-0.0820	0.0134
	Outros	0.0478	0.00715
	Private	0.2521	0.0110
	RNA	0	0
	Unidade Bancárias e Parcerias	-0.0714	0.0206

Quadro C.1 – Continuação da página anterior

Variável	Valores	$\hat{\beta}$	Desvio Padrão
Produto	Protect	0	0
	Ice 3	-0.3353	0.00826
	Protocolos-Ordens	0.0776	0.0118
	Protocolos-Renault	0.3163	0.0253
	Protocolos- Financeiras	0.3415	0.0289
	Protocolos-Barclays	0.4843	0.0462
	Protocolos-FSegurança	0.2152	0.0143
	Protocolos-Outros	0.0696	0.0187
	Protocolos-Funcionários	0.0817	0.0642
	Protocolos- seguros especiais	0.2162	0.0386

Quadro C.1 – Continuação da página anterior

Anexo D

Modelo de Previsão de Anulação do Contrato de Seguro Automóvel em Excel

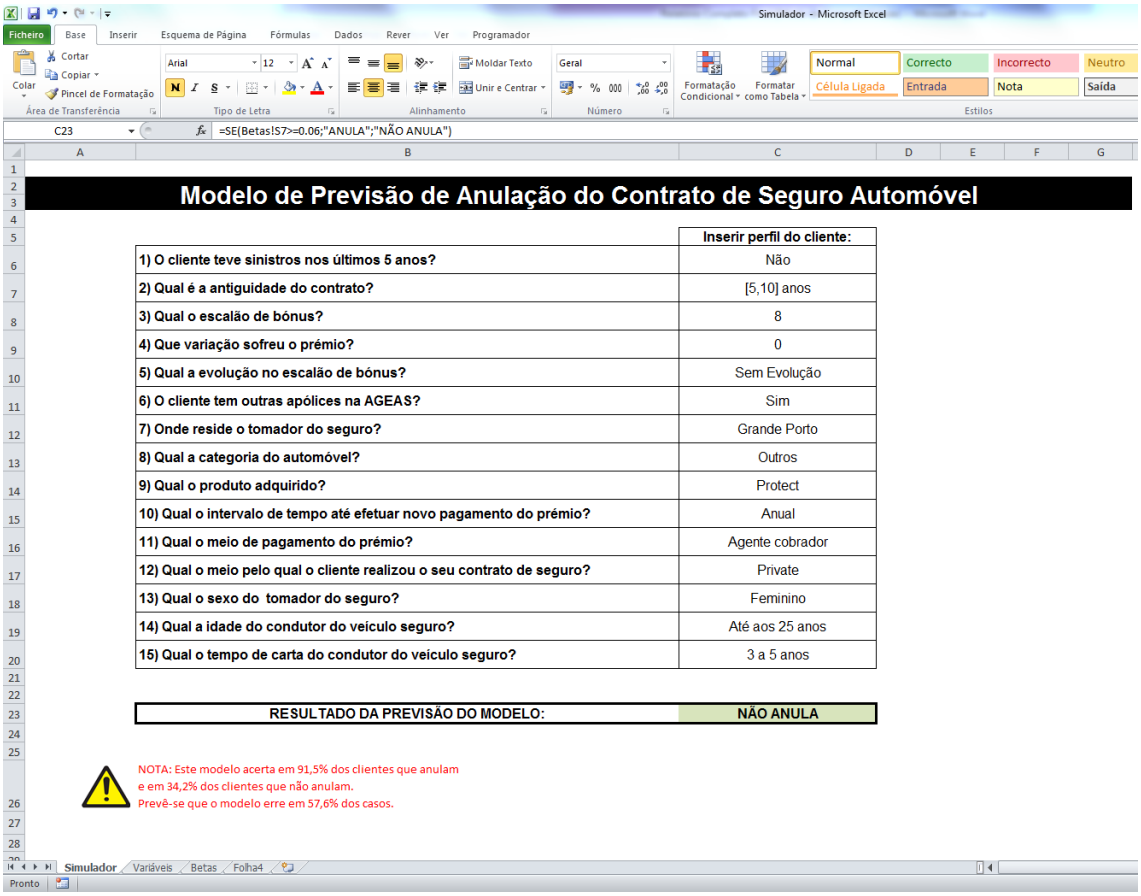
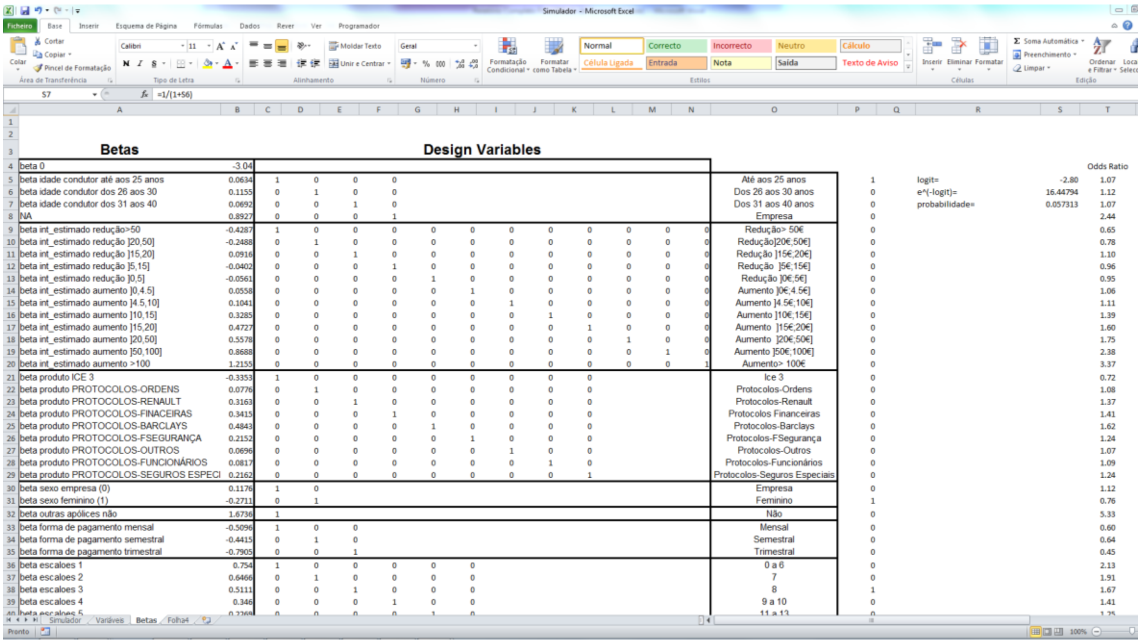


Figura D. 1 - Modelo de Previsão de Anulação do Contrato de Seguro Automóvel aplicado em Excel

O modelo de regressão logística final foi aplicado em excel como se pode ver pela figura D.1 para uma fácil utilização. O resultado final (Cliente anula ou não anula) depende da fórmula inserida na célula C23 que considerando o ponto de corte de 0.06, devolve o resultado " NÃO ANULA" se a célula S7 da folha "Betas" é <0.06 e devolve o resultado "ANULA" caso contrário.

Por sua vez, a célula "S7" da folha "Betas" calcula o valor da resposta do modelo de regressão logística como se pode ver pela figura D.2:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S	T
3	**Betas**						**Design Variables**												
4	beta 0	-3.04																	Odds Ratio
5	beta idade condutor até aos 25 anos	0.0634	1	0	0	0									Até aos 25 anos	1	logit=	-2.80	1.07
6	beta idade condutor dos 26 aos 30	0.1155	0	1	0	0									Dos 26 aos 30 anos	0	e^(-logit)=	16.44794	1.12
7	beta idade condutor dos 31 aos 40	0.0692	0	0	1	0									Dos 31 aos 40 anos	0	probabilidade=	0.057313	1.07
8	NA	0.8927	0	0	0	1									Empresa	0			2.44
9	beta int_estimado redução>50	-0.4287	1	0	0	0	0	0	0	0	0	0	0	0	Redução> 50€	0			0.65
10	beta int_estimado redução]20,50]	-0.2488	0	1	0	0	0	0	0	0	0	0	0	0	Redução]20€;50€]	0			0.78
11	beta int_estimado redução]15,20]	0.0916	0	0	1	0	0	0	0	0	0	0	0	0	Redução]15€;20€]	0			1.10
12	beta int_estimado redução]5,15]	-0.0402	0	0	0	1	0	0	0	0	0	0	0	0	Redução]5€;15€]	0			0.96
13	beta int_estimado redução]0,5]	-0.0561	0	0	0	0	1	0	0	0	0	0	0	0	Redução]0€;5€]	0			0.95
14	beta int_estimado aumento]0,4.5]	0.0558	0	0	0	0	0	1	0	0	0	0	0	0	Aumento]0€;4.5€]	0			1.06
15	beta int_estimado aumento]4.5,10]	0.1041	0	0	0	0	0	0	1	0	0	0	0	0	Aumento]4.5€;10€]	0			1.11
16	beta int_estimado aumento]10,15]	0.3285	0	0	0	0	0	0	0	1	0	0	0	0	Aumento]10€;15€]	0			1.39
17	beta int_estimado aumento]15,20]	0.4727	0	0	0	0	0	0	0	0	1	0	0	0	Aumento]15€;20€]	0			1.60
18	beta int_estimado aumento]20,50]	0.5578	0	0	0	0	0	0	0	0	0	1	0	0	Aumento]20€;50€]	0			1.75
19	beta int_estimado aumento]50,100]	0.8688	0	0	0	0	0	0	0	0	0	0	1	0	Aumento]50€;100€]	0			2.38
20	beta int_estimado aumento >100	1.2155	0	0	0	0	0	0	0	0	0	0	0	1	Aumento> 100€	0			3.37
21	beta produto ICE 3	-0.3353	1	0	0	0	0	0	0	0	0				Ice 3	0			0.72
22	beta produto PROTOCOLOS-ORDENS	0.0776	0	1	0	0	0	0	0	0	0				Protocolos-Ordens	0			1.08
23	beta produto PROTOCOLOS-RENAULT	0.3163	0	0	1	0	0	0	0	0	0				Protocolos-Renault	0			1.37
24	beta produto PROTOCOLOS-FINACEIRAS	0.3415	0	0	0	1	0	0	0	0	0				Protocolos Financeiras	0			1.41
25	beta produto PROTOCOLOS-BARCLAYS	0.4843	0	0	0	0	1	0	0	0	0				Protocolos-Barclays	0			1.62
26	beta produto PROTOCOLOS-FSEGURANÇA	0.2152	0	0	0	0	0	1	0	0	0				Protocolos-FSegurança	0			1.24
27	beta produto PROTOCOLOS-OUTROS	0.0696	0	0	0	0	0	0	1	0	0				Protocolos-Outros	0			1.07
28	beta produto PROTOCOLOS-FUNCIONÁRIOS	0.0817	0	0	0	0	0	0	0	1	0				Protocolos-Funcionários	0			1.09
29	beta produto PROTOCOLOS-SEGUROS ESPECI	0.2162	0	0	0	0	0	0	0	0	1				Protocolos-Seguros Especiais	0			1.24
30	beta sexo empresa (0)	0.1176	1	0											Empresa	0			1.12
31	beta sexo feminino (1)	-0.2711	0	1											Feminino	1			0.76
32	beta outras apólices não	1.6736	1												Não	0			5.33
33	beta forma de pagamento mensal	-0.5096	1	0	0										Mensal	0			0.60
34	beta forma de pagamento semestral	-0.4415	0	1	0										Semestral	0			0.64
35	beta forma de pagamento trimestral	-0.7905	0	0	1										Trimestral	0			0.45
36	beta escaloes 1	0.754	1	0	0	0	0	0							0 a 6	0			2.13
37	beta escaloes 2	0.6466	0	1	0	0	0	0							7	0			1.91
38	beta escaloes 3	0.5111	0	0	1	0	0	0							8	1			1.67
39	beta escaloes 4	0.346	0	0	0	1	0	0							9 a 10	0			1.41
40	beta escaloes 5	0.2269	0	0	0	0	1	0							11 a 13	0			1.25

Figura D. 2 - Folha Betas do Modelo de Previsão de Anulação do Contrato de Seguro Automóvel aplicado em Excel